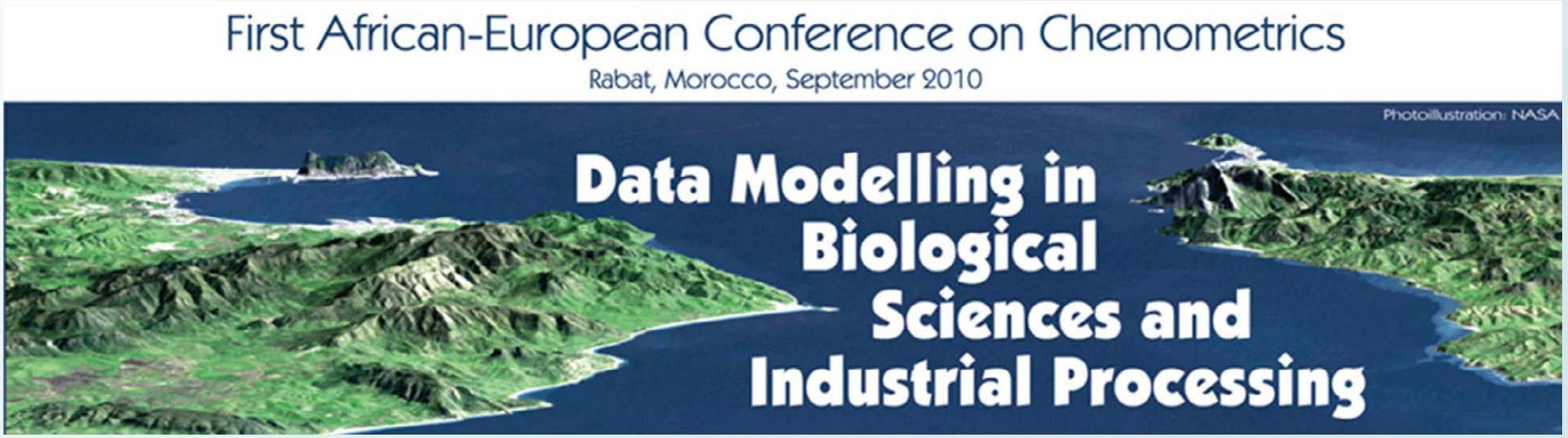
Calibration transfer techniques and spectra control in the framework of breeding programs

Barreiro, P, Moya-González, A
Technical University of Madrid



Why this topic...

Breeding processes take the advantage of the NIRS being non destructive, while

NIRS takes profit of the large range of variation of quality parameters within the former, which makes them very suitable for model calibration

Justification of breeding programs

Crop relevance

- Elite varieties non available

Genetic constraints

- Breeding program for some quality attributes
  - Low heritability
  - MAS no possible
  - High variability

Selected breeding program

- Bulk selection
  - High number of individuals→ Extension over time

Let's focus

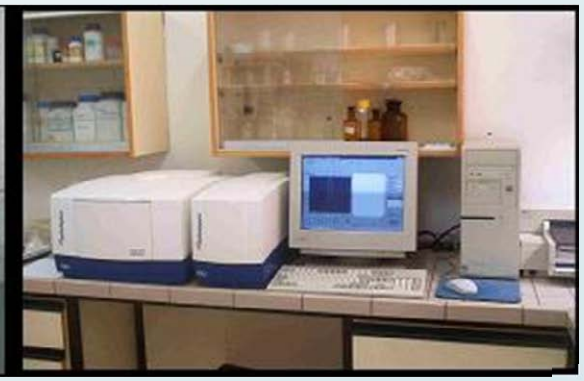
- Olive breeding at UCO-IFAPA (Córdoba)
- Onion breeding at Agrotécnica (Ext remadura)

- Olive breeding at IFAPA (Córdoba)

- Onion breeding at Agrotécnica (Ext remadura)

How destructive can be destructive analysis?

Oven drying

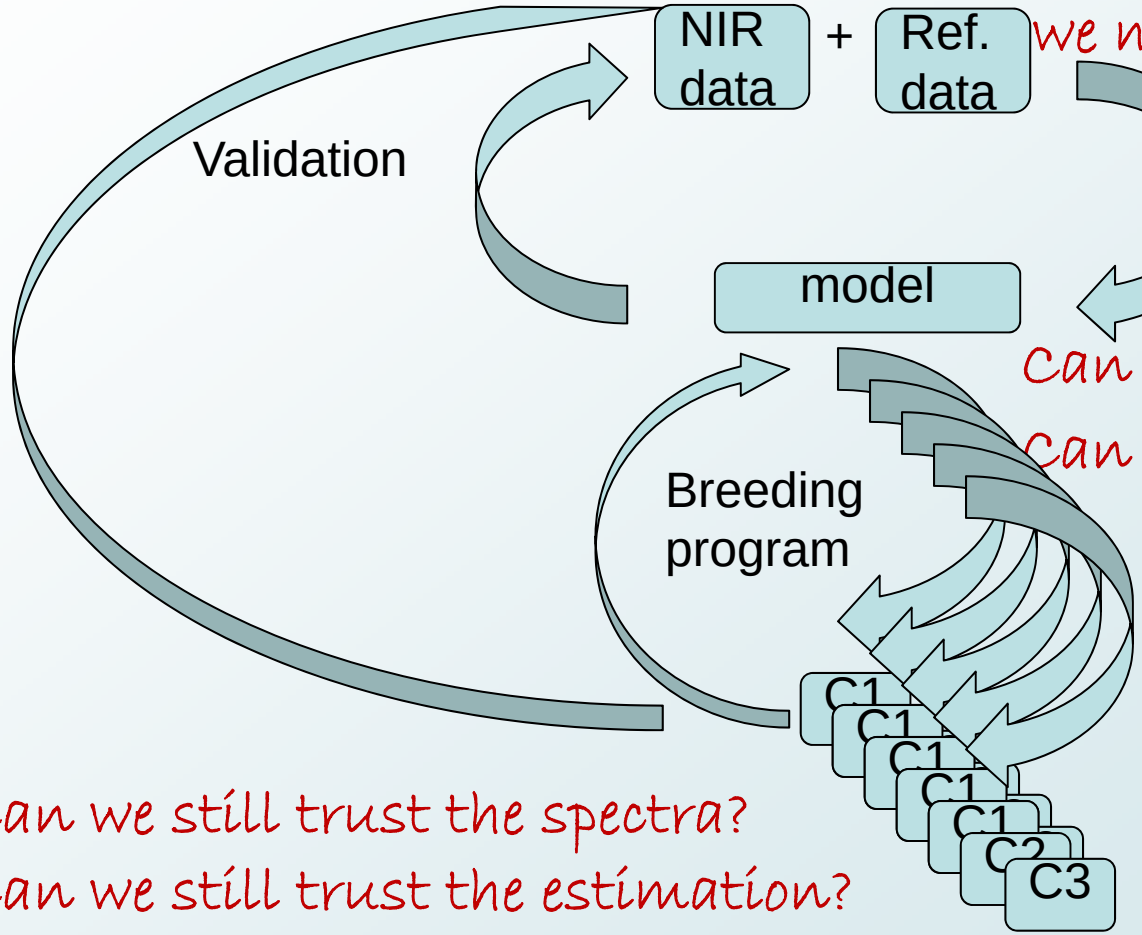


Bruker.
The minispec

UNIVERSIDAD D CÓRDOBA

A little bit of System Thinking

Can we avoid calibration when we modify the equipment?



Can we trust the spectra?

Can we trust the estimation?

Can we still trust the spectra?

Can we still trust the estimation?

thus...

Can you picture yourself changing your instrument several times during the process?

...

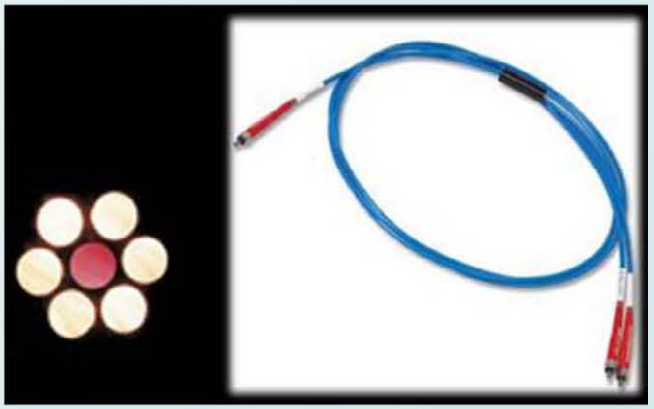
What about the effect of the agroevolution?

So the Calibration Transfer Concept is ...

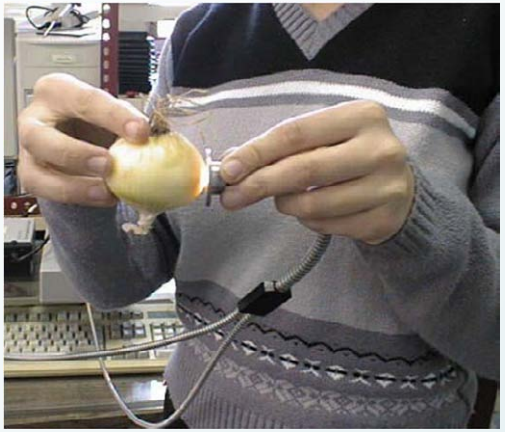
To run all the time, contrawise all other variations, trying to remain in the same position

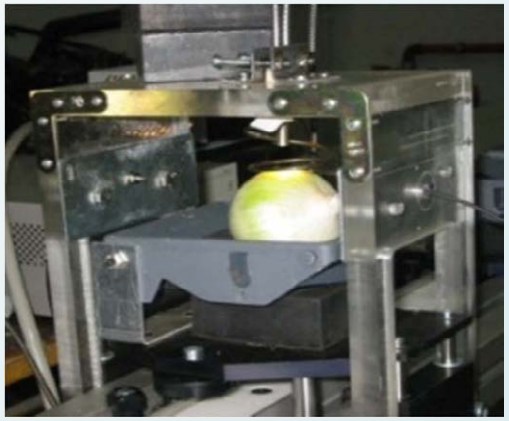
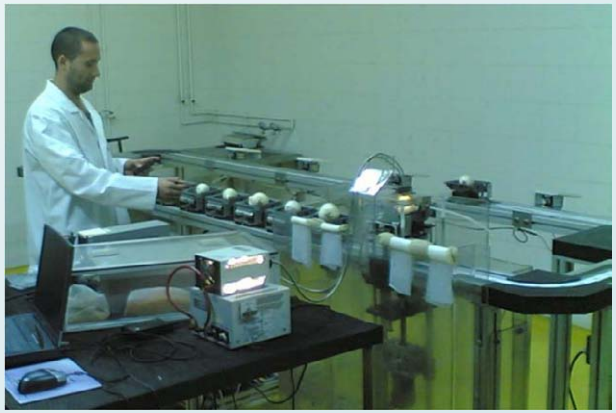
STANDARD LABORATORY NIRS

Our potable equipment for olives

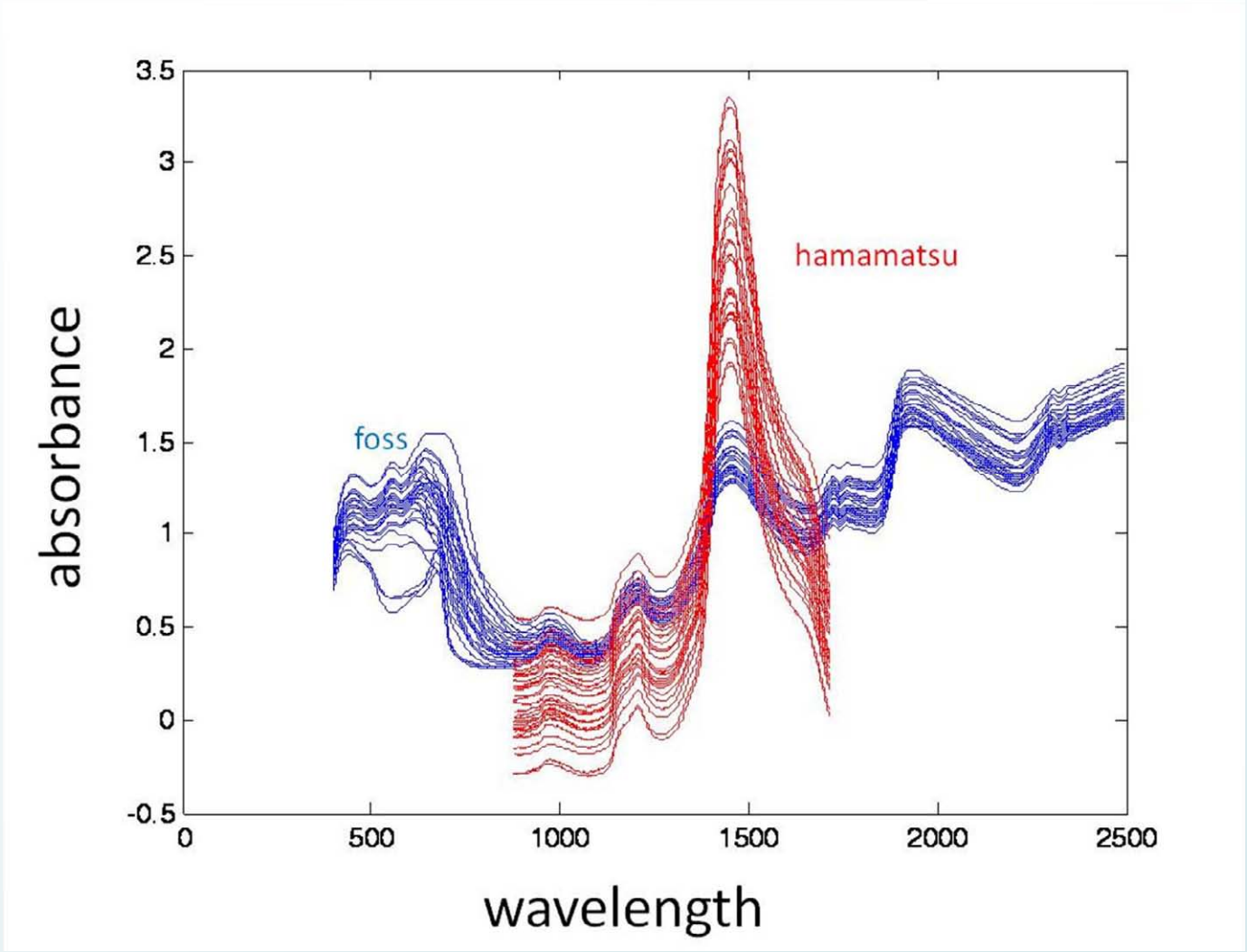


our automated grading line for NIRS based classification of onion bulbs



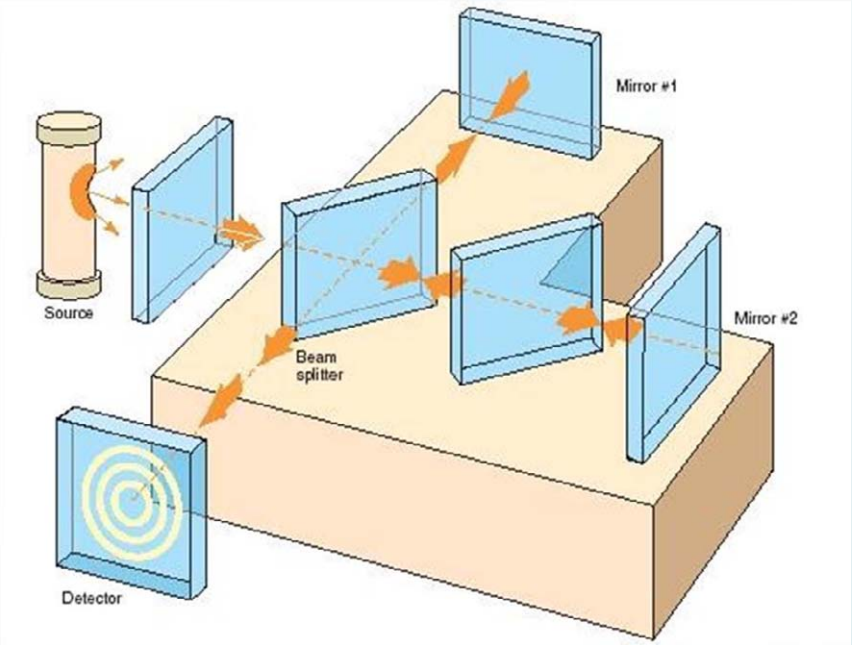



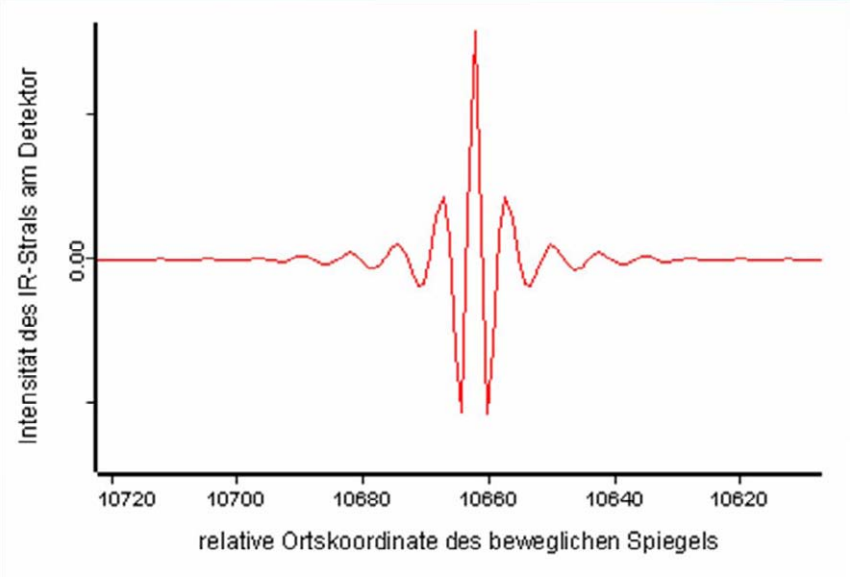
An example of 2 instrument spectra



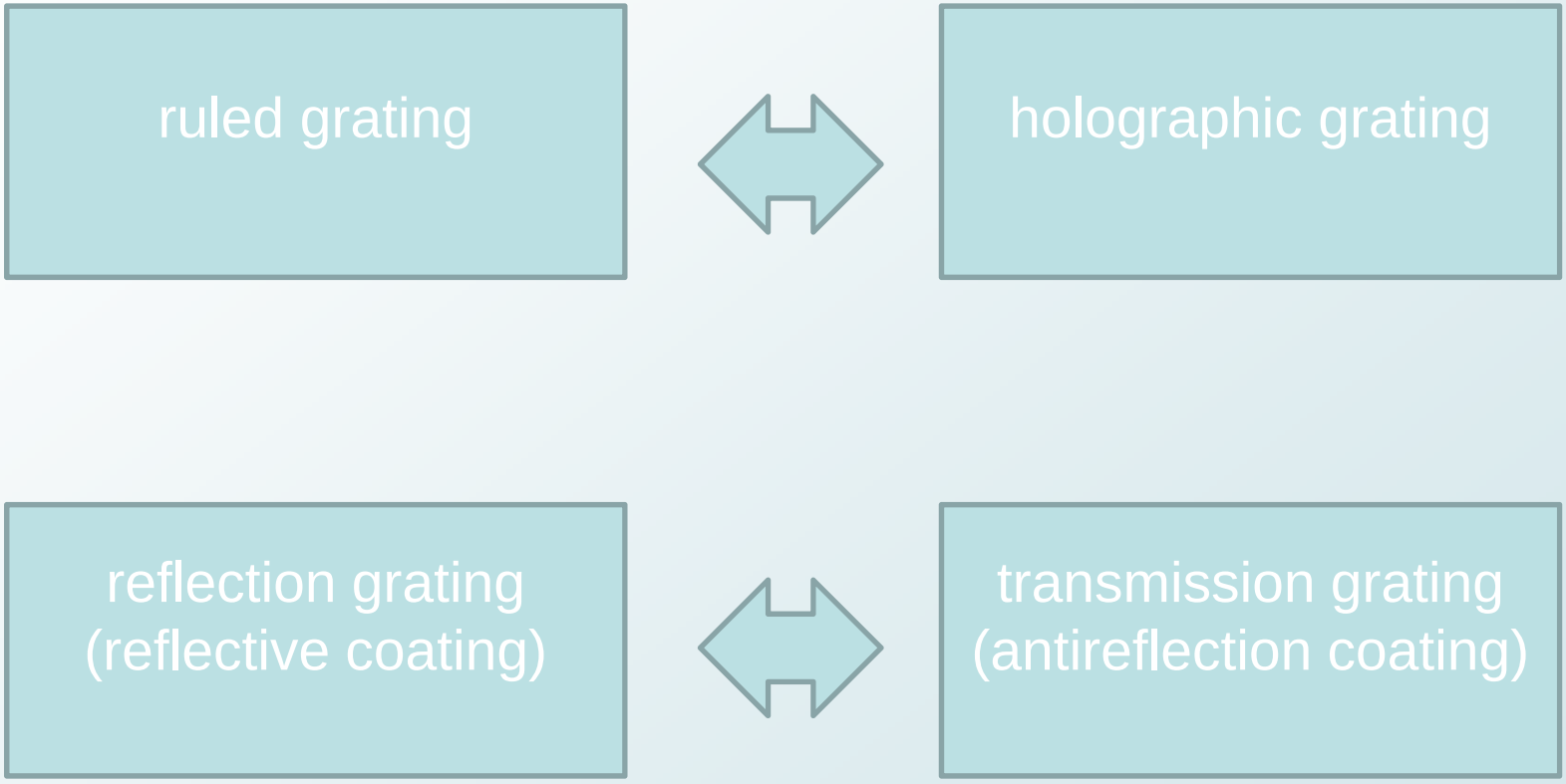
How deep we need to go into the instrument Knowledge?

michelson interferometer



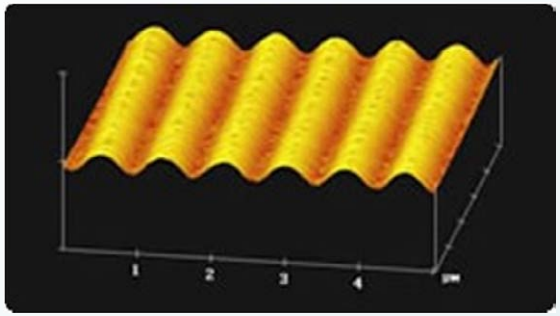


Improvements in diffraction Gratings

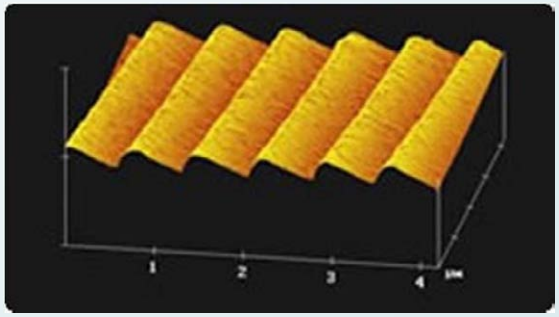


Diffraction Gratings Profile

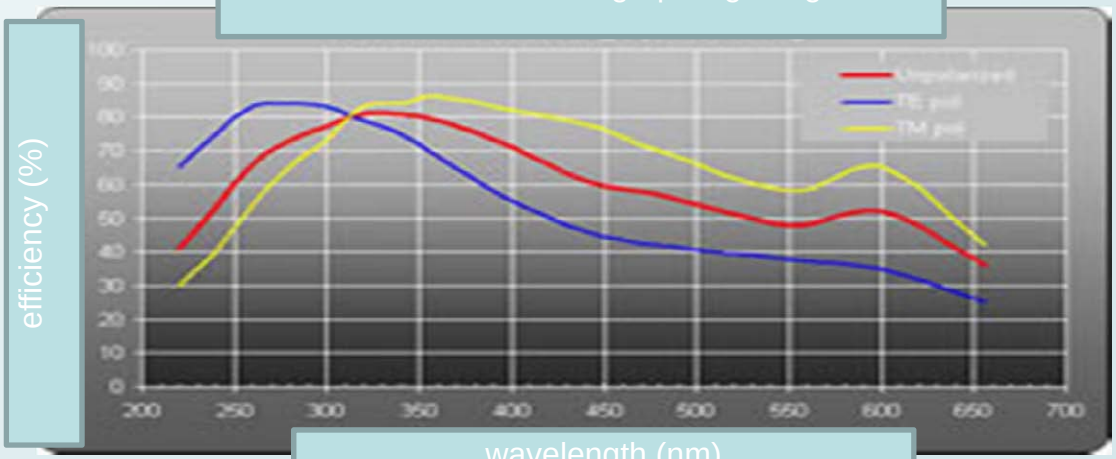
sinusoidal profile



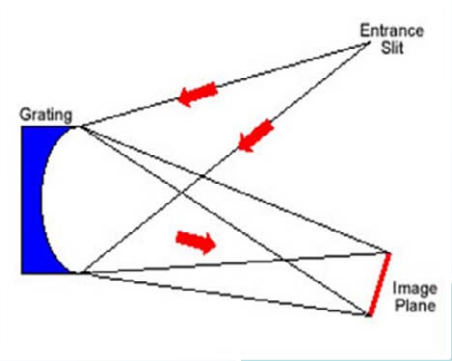
blazed profile



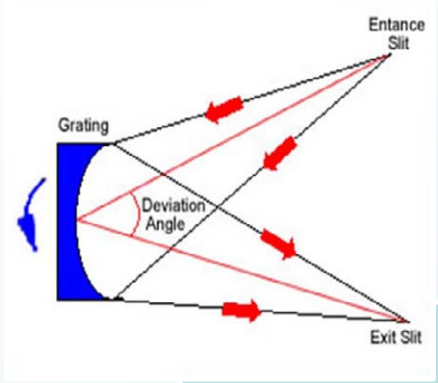
325 nm blazed hoographic grating



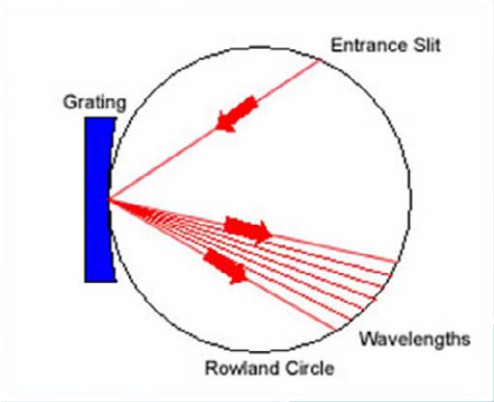
Concave Gratings



- Aberration Corrected (Flat Field Imaging) Concave Grating

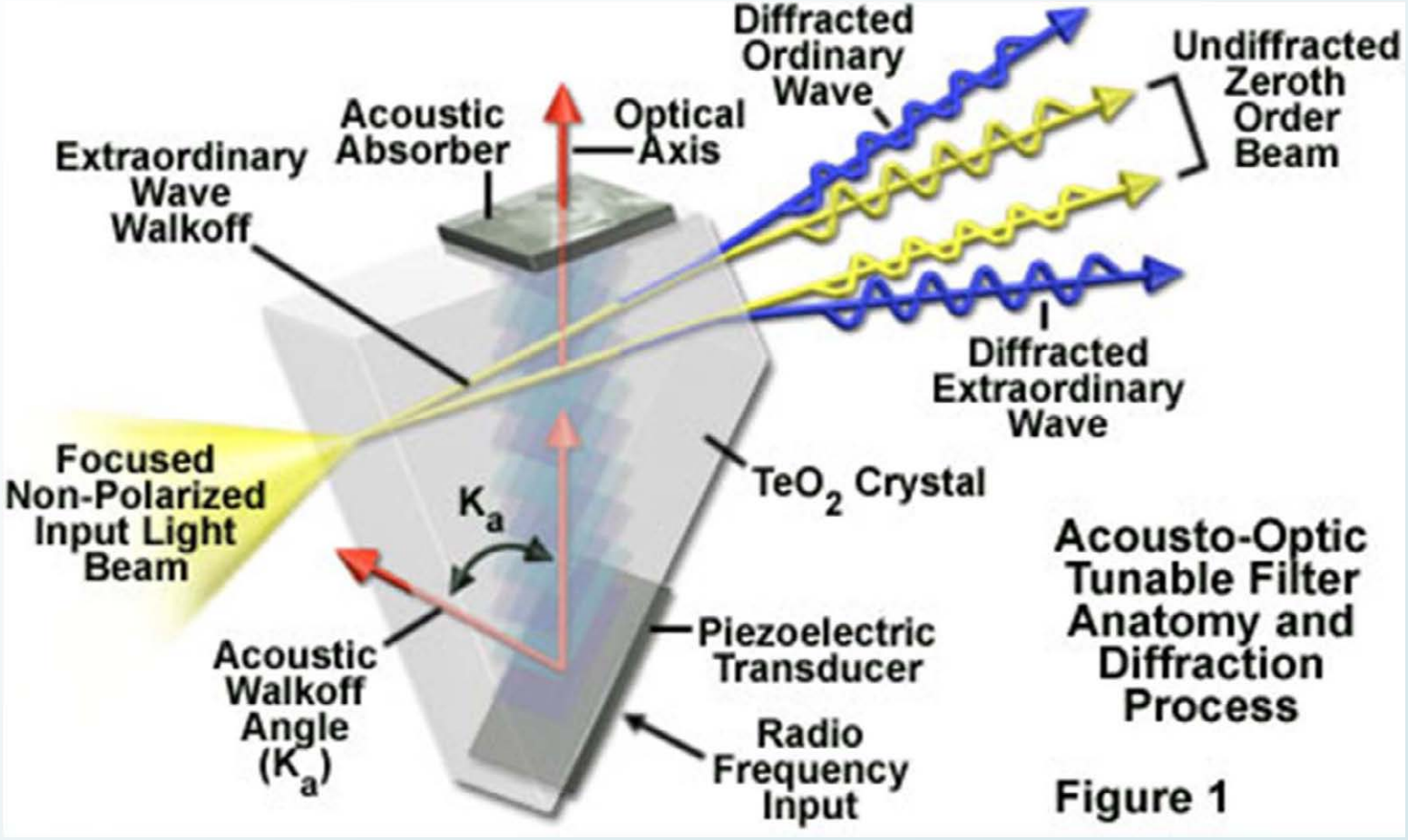


- Constant Deviation Monochromator Concave Grating



- Rowland Type Concave Grating

AOTF principles



The very basics of model calibration

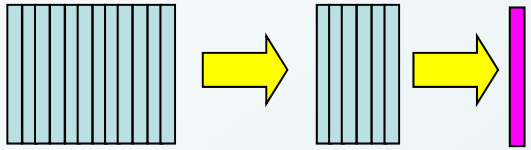
- The methods:
  - MLR stepwise
  - PCR
  - PLS
  - PLS + wave selection
- The quality of estimation:
  - R2
  - SEC, SEP
  - PDR=SEP/STD

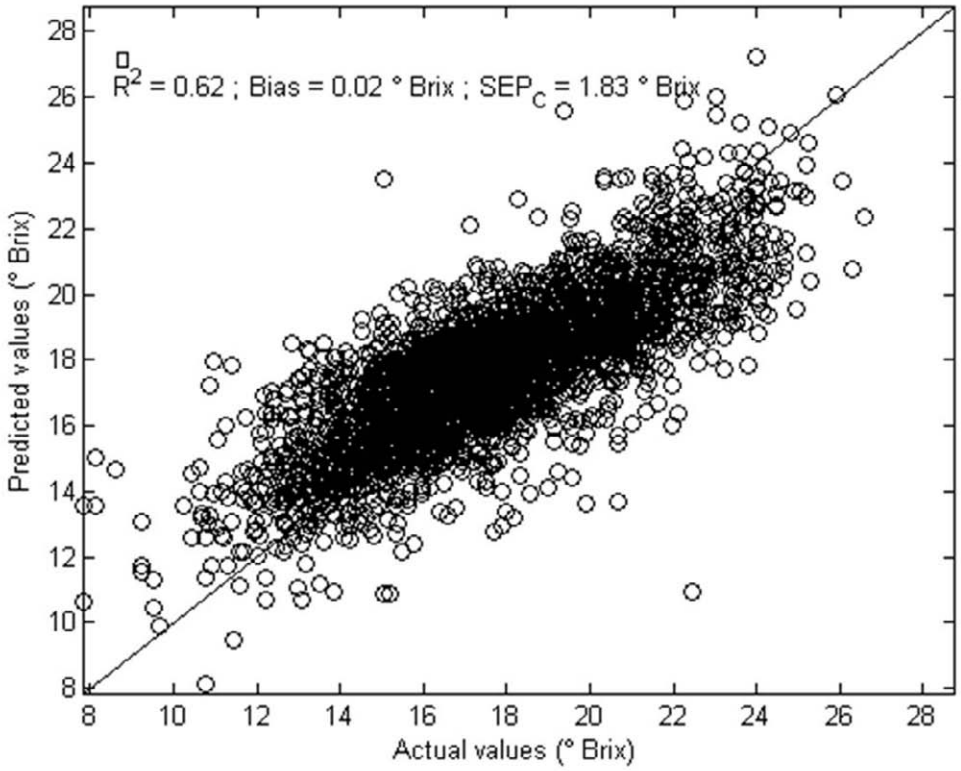
MLR model calibrated from at-line spectra

at-line & on-line validation

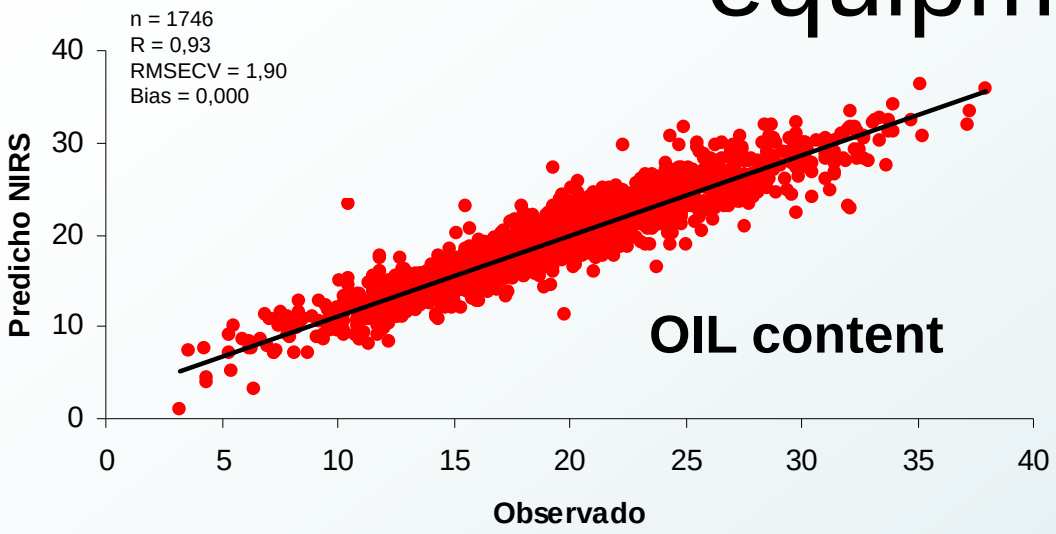
$$SSC = 14{,}54 + 1{,}89\lambda_{894} - 27{,}08\lambda_{907} + 4{,}09\lambda_{947} - 1{,}54\lambda_{1000} - 5{,}19\lambda_{1187} + 5{,}20\lambda_{1284}$$

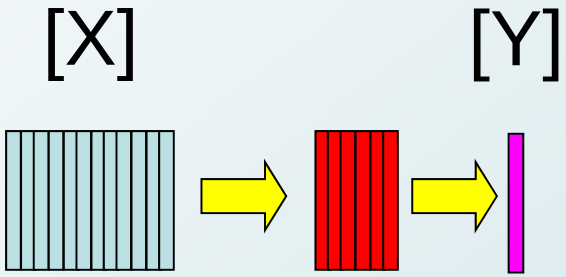
[X] [Y]



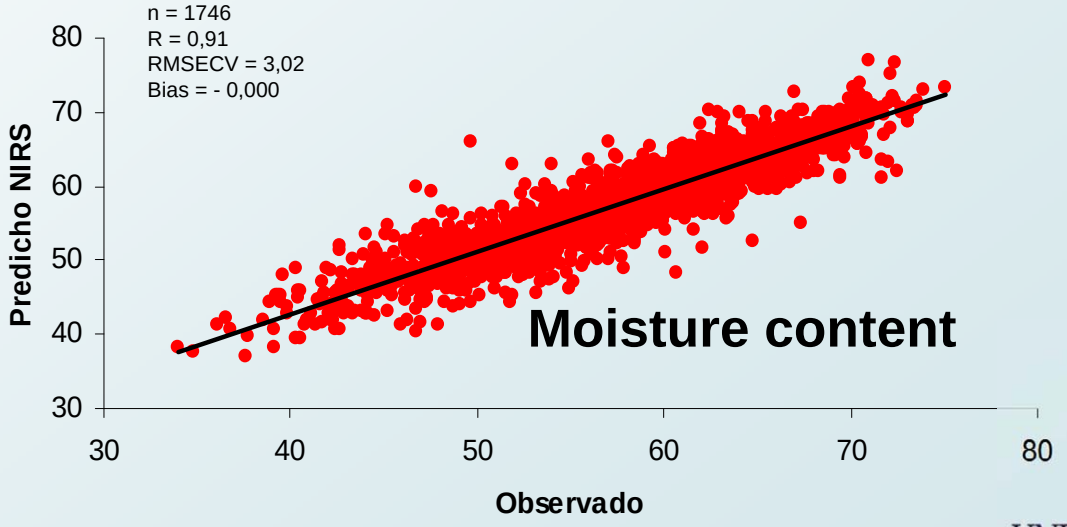


Global PLS models with laboratory equipment





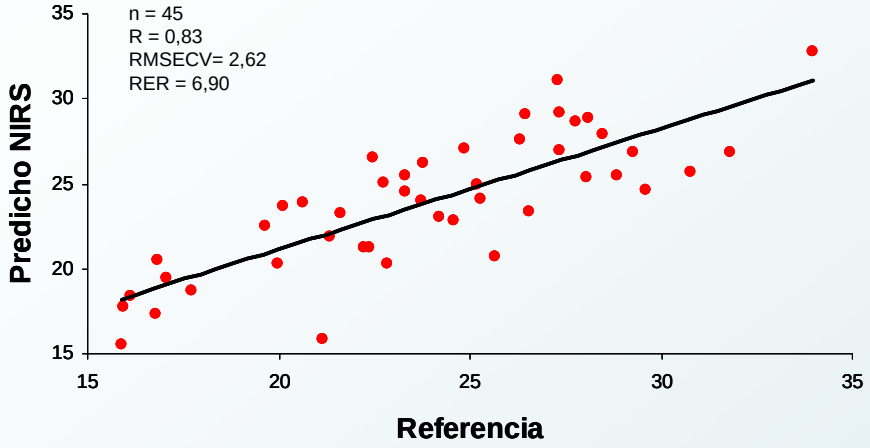
Ecuación de calibración global en % de humedad



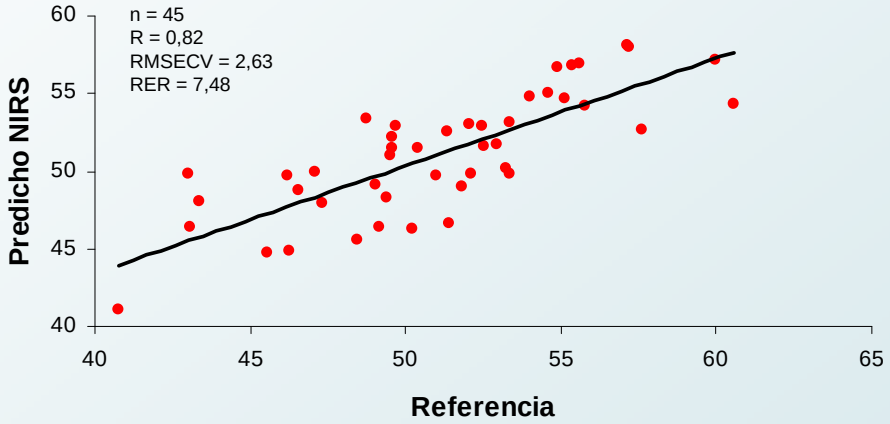
UNIVERSIDAD D CÓRDOBA

Global PLS models for portable equipment

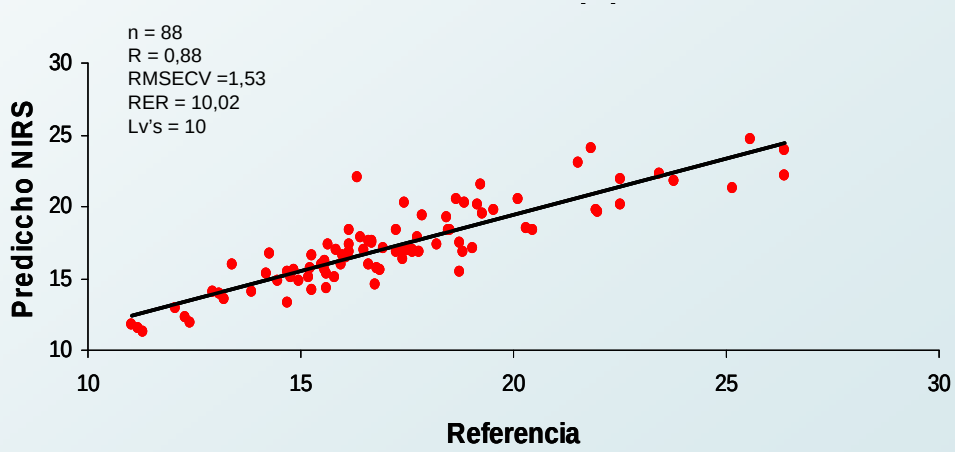
OIL content: 2007/08



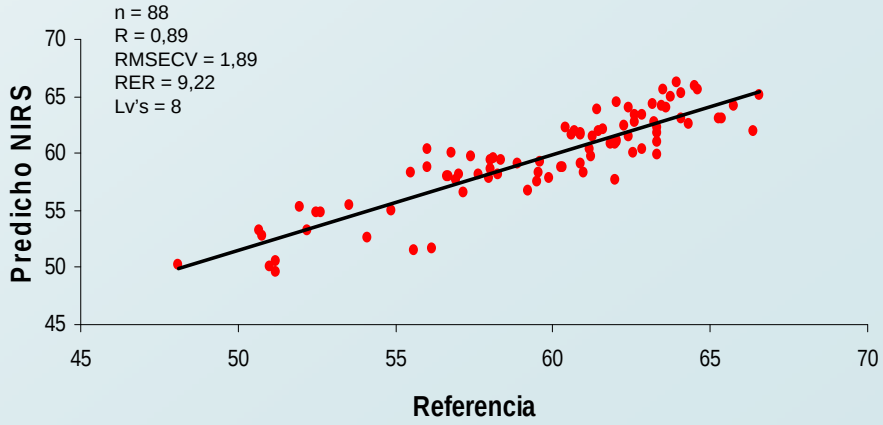
Ecuación de calibración en % de humedad



Moisture content: 2008/09



Ecuación de calibración en % de humedad



Routine PLS with external validation

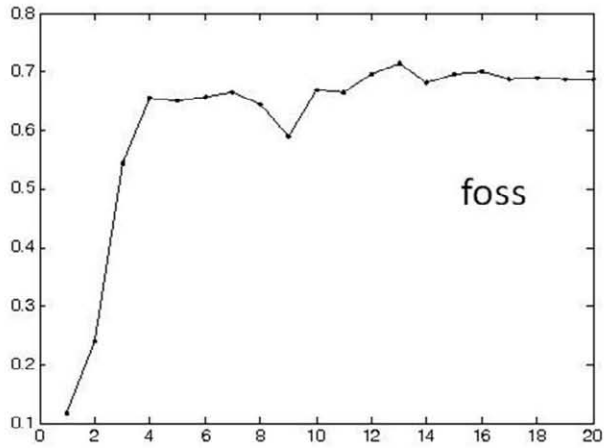
RESUL

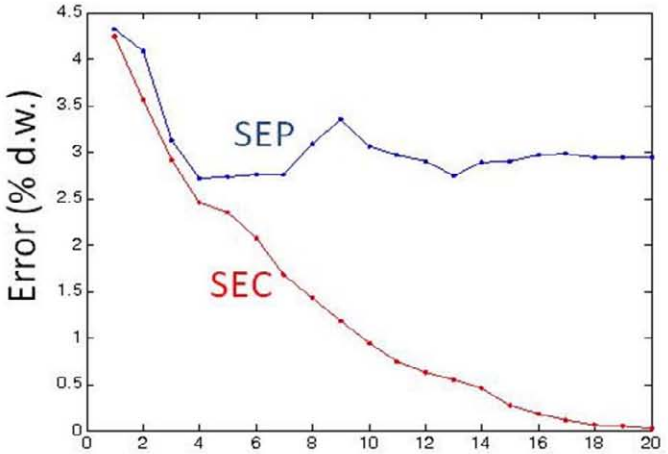
Each

{pred

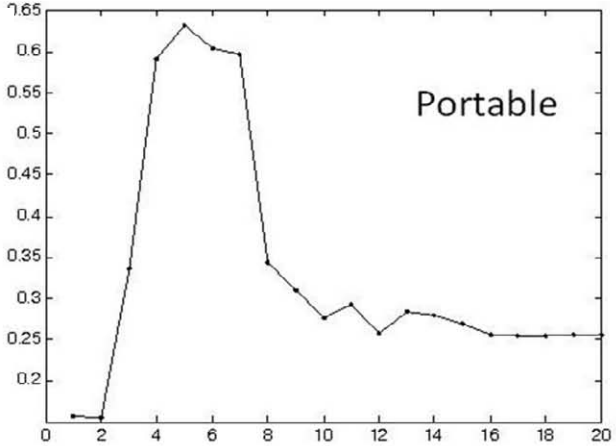
[full_r

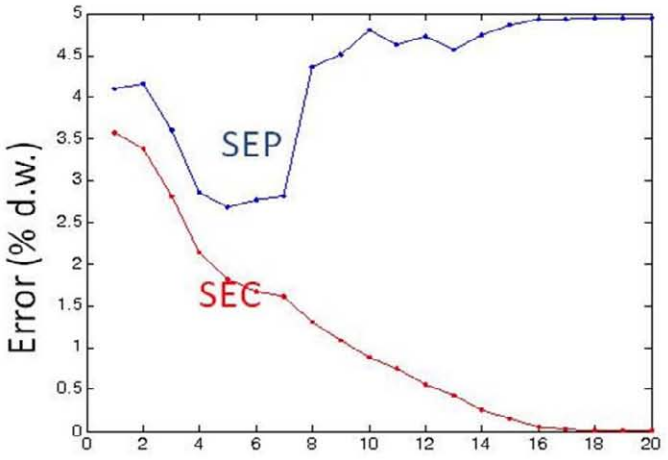
R2



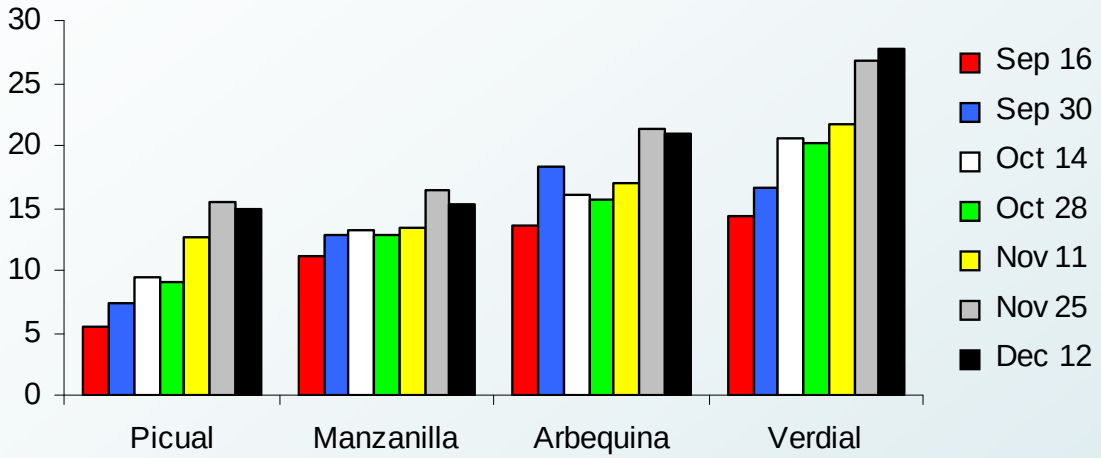


R2

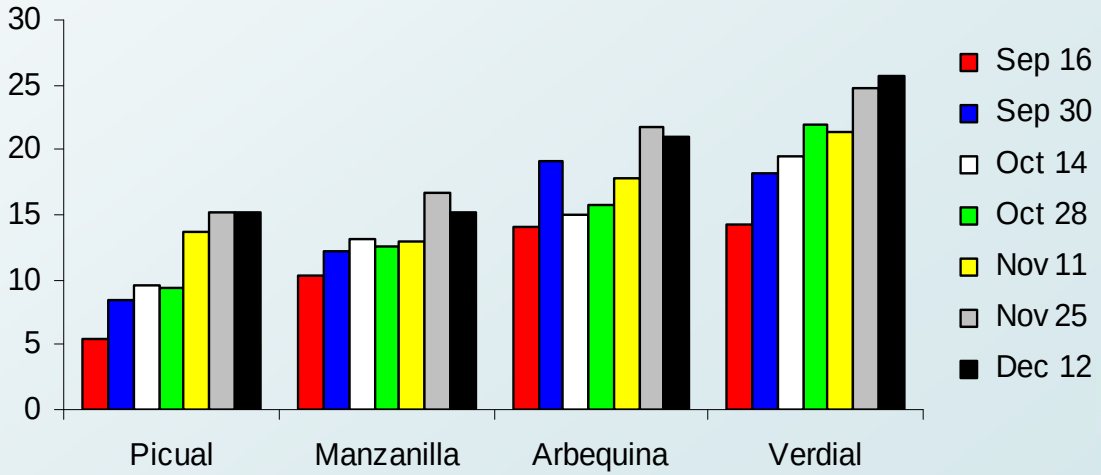




Oil content by reference NMR



Oil content by reference NIRS



UNIVERSIDAD D CÓRDOBA

The loss of estimation performance (in oil)

UNIVERSIDAD D CÓRDOBA

R2	2004	2005	2006	2007
2004	0.905	0.781	0.849	0.692
2005	0.856	0.914	0.829	0.770
2006	0.857	0.865	0.883	0.739
2007	0.838	0.724	0.787	0.885

RMSEP	2004	2005	2006	2007
2004	1.222	2.464	4.563	7.165
2005	2.013	1.393	5.143	8.577
2006	5.966	4.478	1.694	3.141
2007	9.796	7.434	4.504	1.601

The loss of estimation performance (in moisture)

UNIVERSIDAD D CÓRDOBA

R2	2004	2005	2006	2007
2004	0.949	0.882	0.876	0.538
2005	0.896	0.917	0.889	0.571
2006	0.910	0.898	0.913	0.668
2007	0.890	0.777	0.803	0.869

RMSEP	2004	2005	2006	2007
2004	1.578	2.469	4.139	5.534
2005	3.717	1.980	3.940	5.002
2006	3.091	3.436	2.070	3.993
2007	2.743	3.413	3.200	1.928

robustness

It refers to the stability of a multivariate model with regards to perturbations from the average standard working conditions:

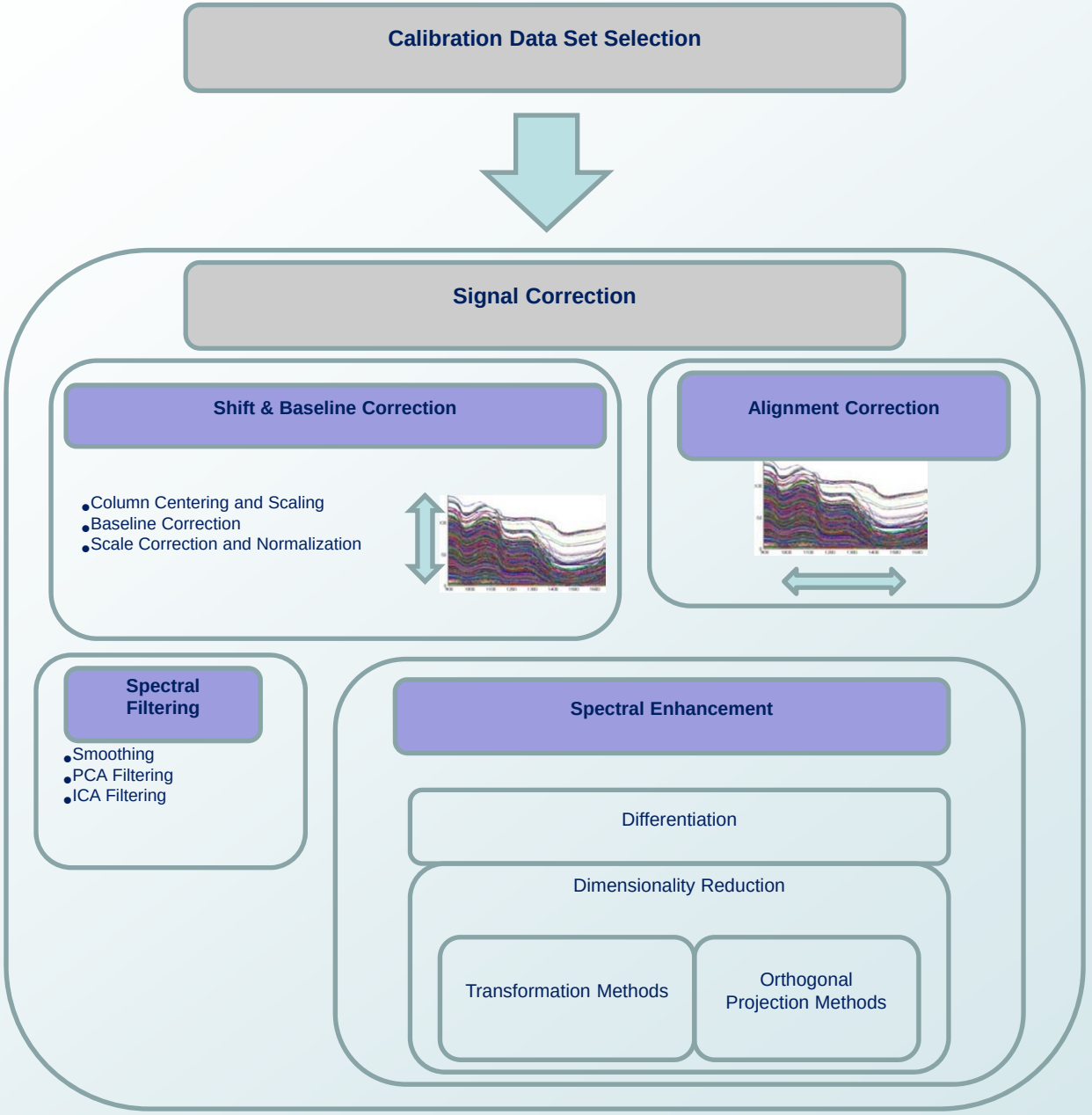
- External robustness
- Internal robustness

Robust modeling seeks for the independence of estimation from external factors without increasing routine analysis

$$\left[\hat{\mathbf{Y}}\right]_{n\times 1} = \left[\mathbf{X}_{n\times m}\right] \times \left[\mathbf{b}\right]_{(m)\times 1}$$

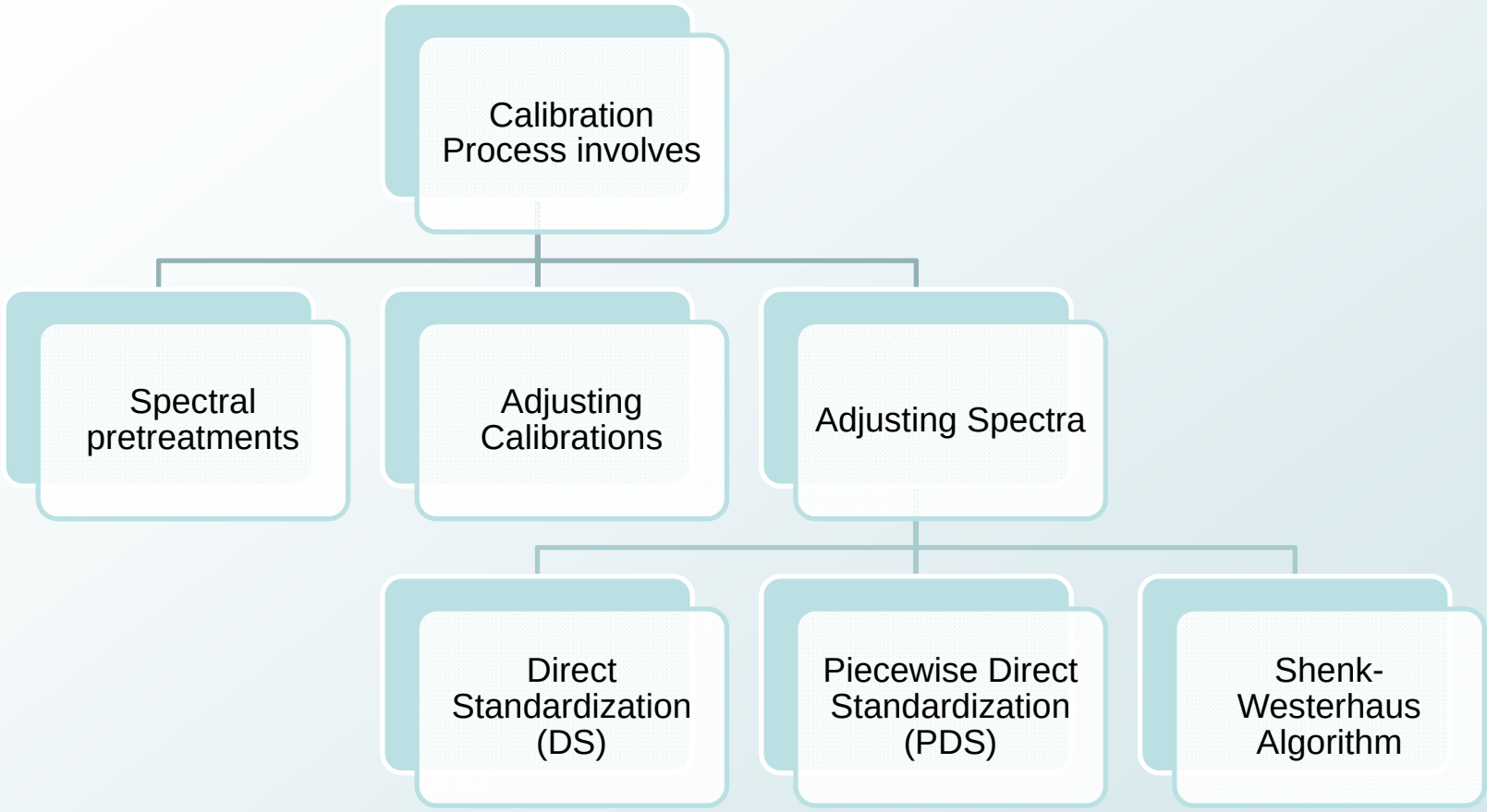
$$\delta y_i = y_i - \hat{y}_i = \left\|\delta x_i\right\| \times \left\|b\right\| \times \cos\left(\delta x_i \times b\right)$$

Zeaitier et al., 2005


Calibration Data Set Selection
Signal Correction
Shift & Baseline Correction
Column Centering and Scaling
Baseline Correction
Scale Correction and Normalization
Alignment Correction
Spectral Filtering
Smoothing
PCA Filtering
ICA Filtering
Spectral Enhancement
Differentiation
Dimensionality Reduction
Transformation Methods
Orthogonal Projection Methods

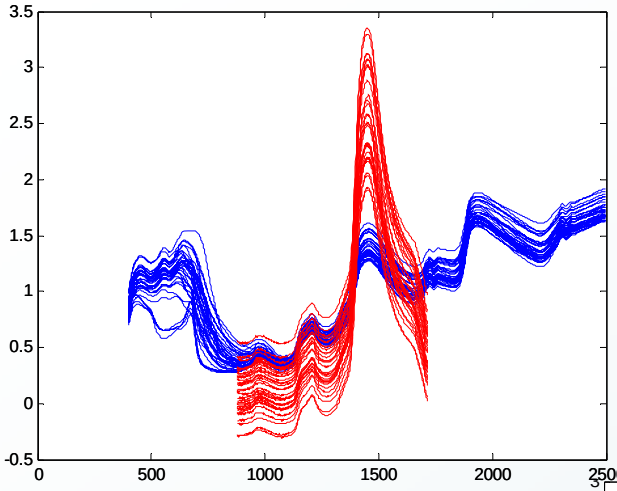
Methods for Calibration Data Set Selection

METHOD	DESCRIPTION	REFERENCE
Random subset selection algorithm	Selection from a random generated index.	(Zeaiter, Rutledge et al. 2009)
Based on ranking y values	Selection from the ranking of the variable of interest values.	(Zeaiter, Rutledge et al. 2009)
Kennard and Stone algorithm	Selects the set of samples that covers the overall spectral experimental domain based on their distance from each other independent of y.	(Kennard and Stone 1969)
Federov algorithm	Selects from a large database those calibration samples that optimally span the domain of interest based on the optimality criterion.	(Pukelsheim 1993)
Duplex algorithm	Variant of the K&S algorithm that makes an alternative selection of the calibration data set and test sets.	(Snee 1977)
Based on cluster analysis	Iteration procedure that selects the samples that are the furthest from the center of each cluster.	(Isaksson and Næs 1990)
Based on factorial analysis	Principal component transformation and selection according to the Mahalanobis distances.	(Puchwein 1988)

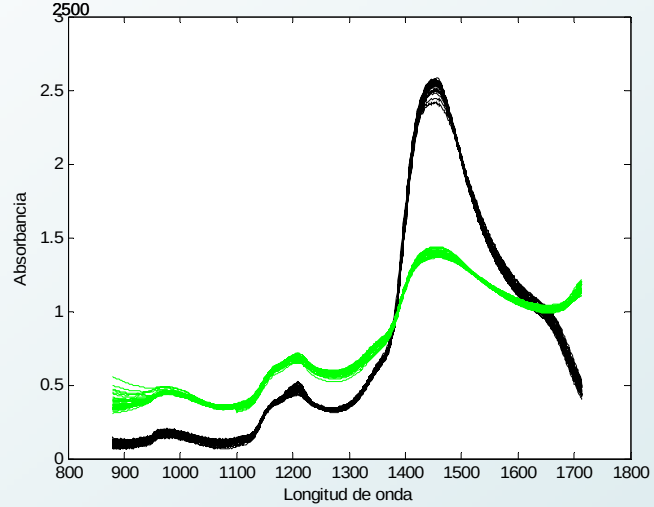

Calibration Process involves
Spectral pretreatments
Adjusting Calibrations
Adjusting Spectra
Direct Standardization (DS)
Piecewise Direct Standardization (PDS)
Shenk-Westerhaus Algorithm

Scale Correction and Normalization

METHOD	DESCRIPTION	REFERENCE
Mean scaling	Each point in the signal is divided by the mean value of that signal.	(Zeaiter, Rutledge et al. 2009)
Maximum scaling	each point in the signal is divided by the maximum value of that signal.	(Zeaiter, Rutledge et al. 2009)
Range scaling	Each point in the signal is divided by the difference between the values at two defined data points.	(Zeaiter, Rutledge et al. 2009)
MinMax scaling	Maximum and minimum values in each signal are set to particular values	(Zeaiter, Rutledge et al. 2009)
Logarithmic scaling	The influence of extreme differences in values of variables can be reduced by applying log scaling	(Zeaiter, Rutledge et al. 2009)
Standard normal variates (SNV) transformation	Each spectrum is centered and then scaled by dividing by its standard deviation	(Barnes, Dhanoa et al. 1989; Barnes, Dhanoa et al. 1993)
SNV-Detrend	After SNV for correction of the linear baseline shift and global signal intensity variations, Detrend corrects any curvilinear trend in the signal baseline by adjusting with a degree 2 polynomial	(Barnes, Dhanoa et al. 1989; Barnes, Dhanoa et al. 1993)
Robust normal variate (RNV) transformation	transformation that modifies SNV by using the percentile instead of the mean	(Guo, Wu et al. 1999)
Multiplicative scatter correction (MSC)	Each individual spectrum is shifted and rotated so that it fits as closely as possible to the chosen reference spectrum for the removal of the diffusion spectrum	(Isaksson and Næs 1988)
Extended multiplicative scatter correction (EMSC)	Removal of the diffusion spectrum using the spectra of analytes and interference effects to compute the correction factors	(Martens and Stark 1991)
Spectral interference subtraction (SIS)	elimination of interferences with known spectral effects	(Martens and Stark 1991)

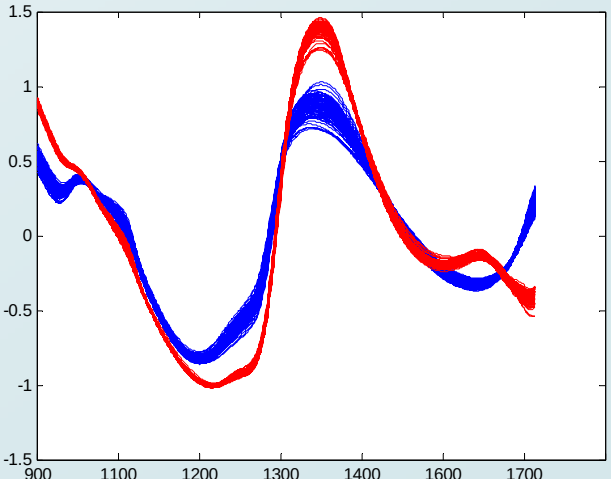


Without pre-peocessing



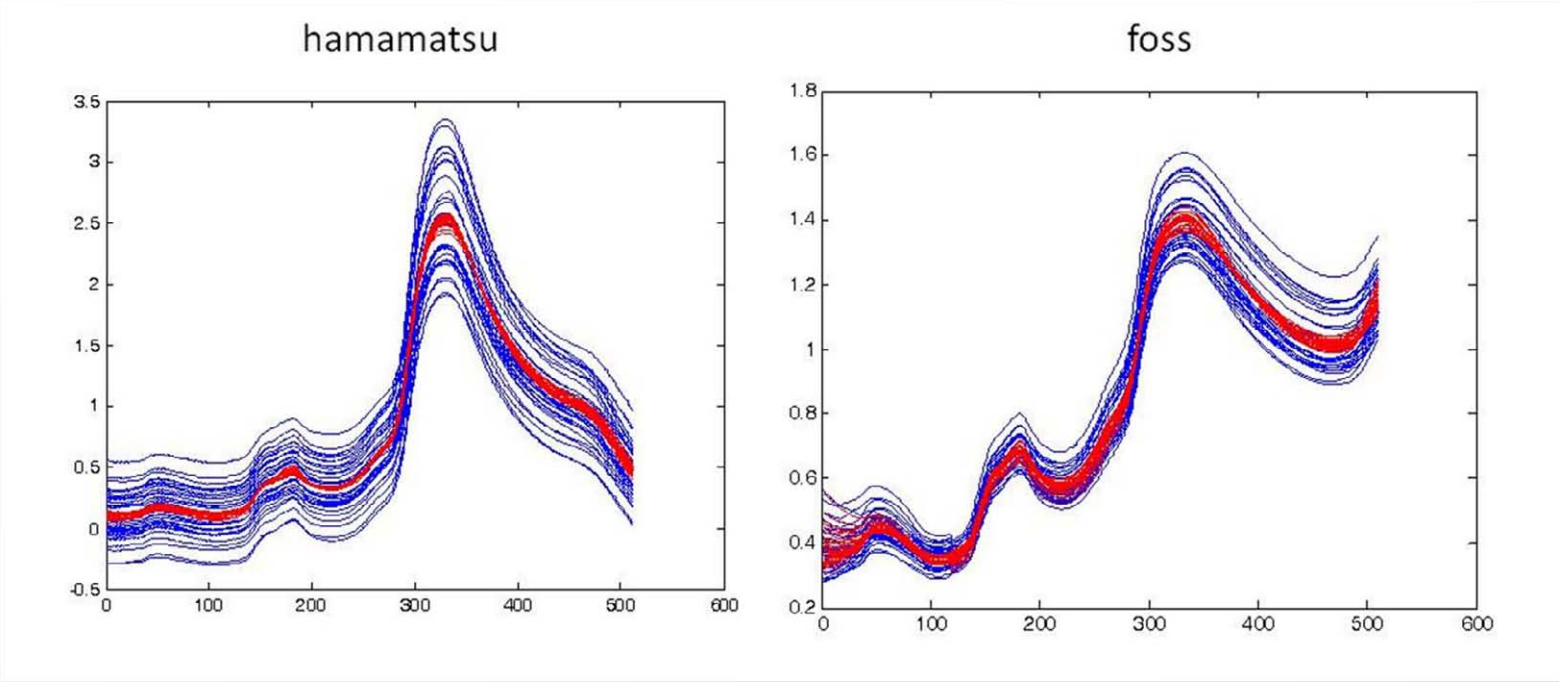
MSC

Savtizky-Golay + SNV + Detrend

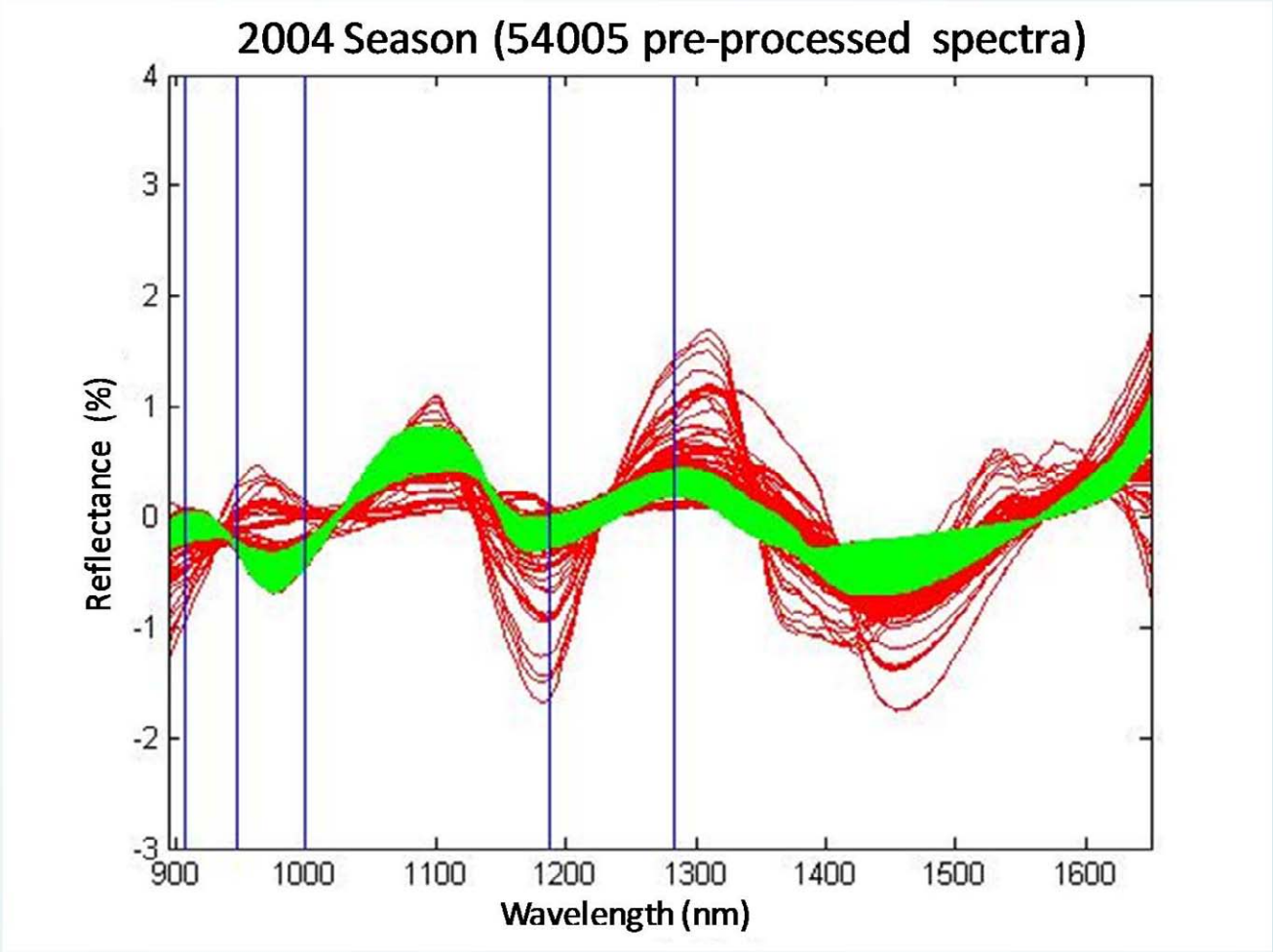


Multiplicative Scatter Correction

Xnew1=msc(X1)
TODO_msc=msc_allseasons(TODO,nspec)



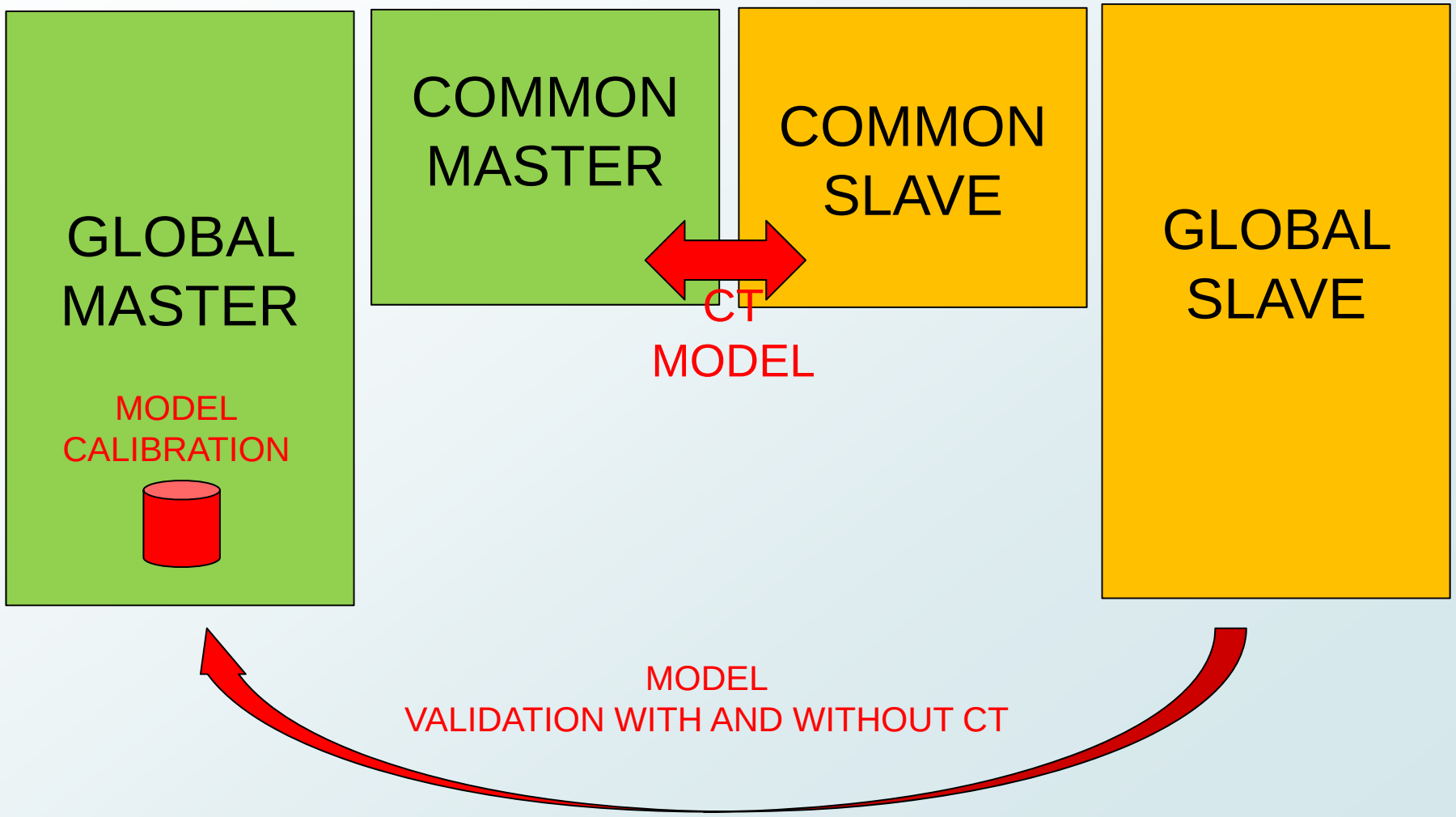
SNV + Detrending



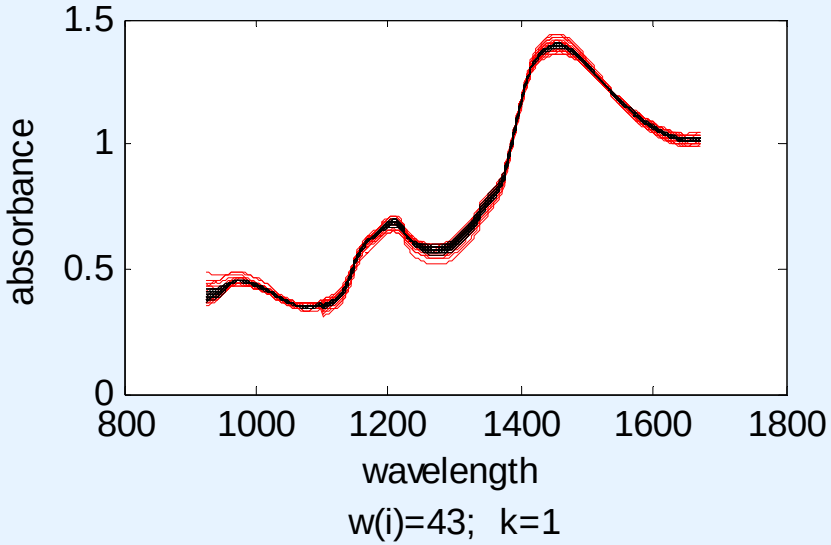
Calibration Transfer Methods

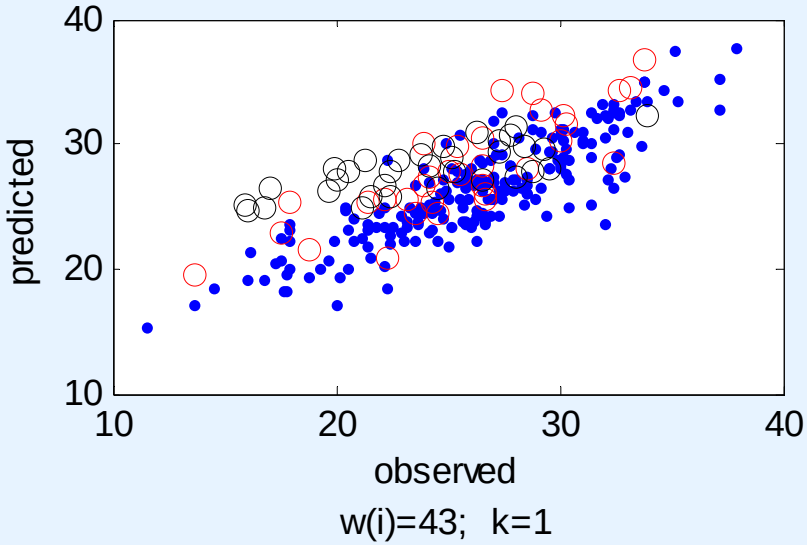
METHOD	DESCRIPTION	REFERENCE
Calibrations adjustment	Slope and bias correction between different instrument measurements.	(Fearn 2001)
Direct Standardization (DS)	Transformation of the original spectra matrix by an F matrix where all the elements can be non cero .	(Wang, Veltkamp et al. 1991)
Piecewise Direct Standardization (PDS)	Transformation of the original spectra matrix by an F matrix where all the non cero elements were distributed around the main diagonal.	(Wang, Veltkamp et al. 1991; Wang, Dean et al. 1995)
Shenk-Westerhaus algorithm	Previous correction for the horizontal displacement and later spectra transformation in the form log(1/R) by means os a diagonal matrix.	(Shenk, Westerhaus et al. 1985; Bouveresse, Massart et al. 1994)

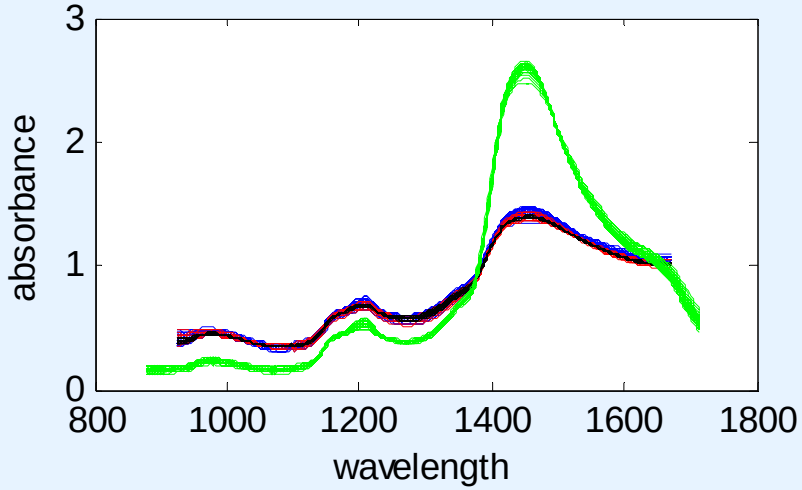
Data Configuration for CT verification

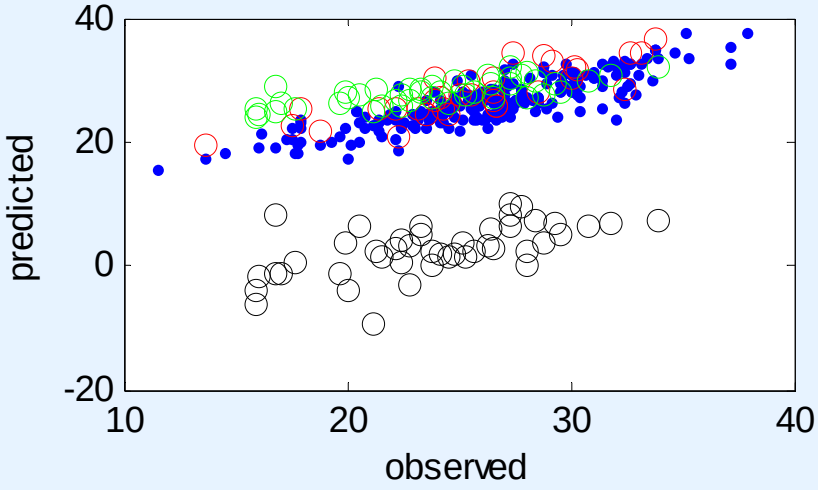


Piecewise Direct Standarization

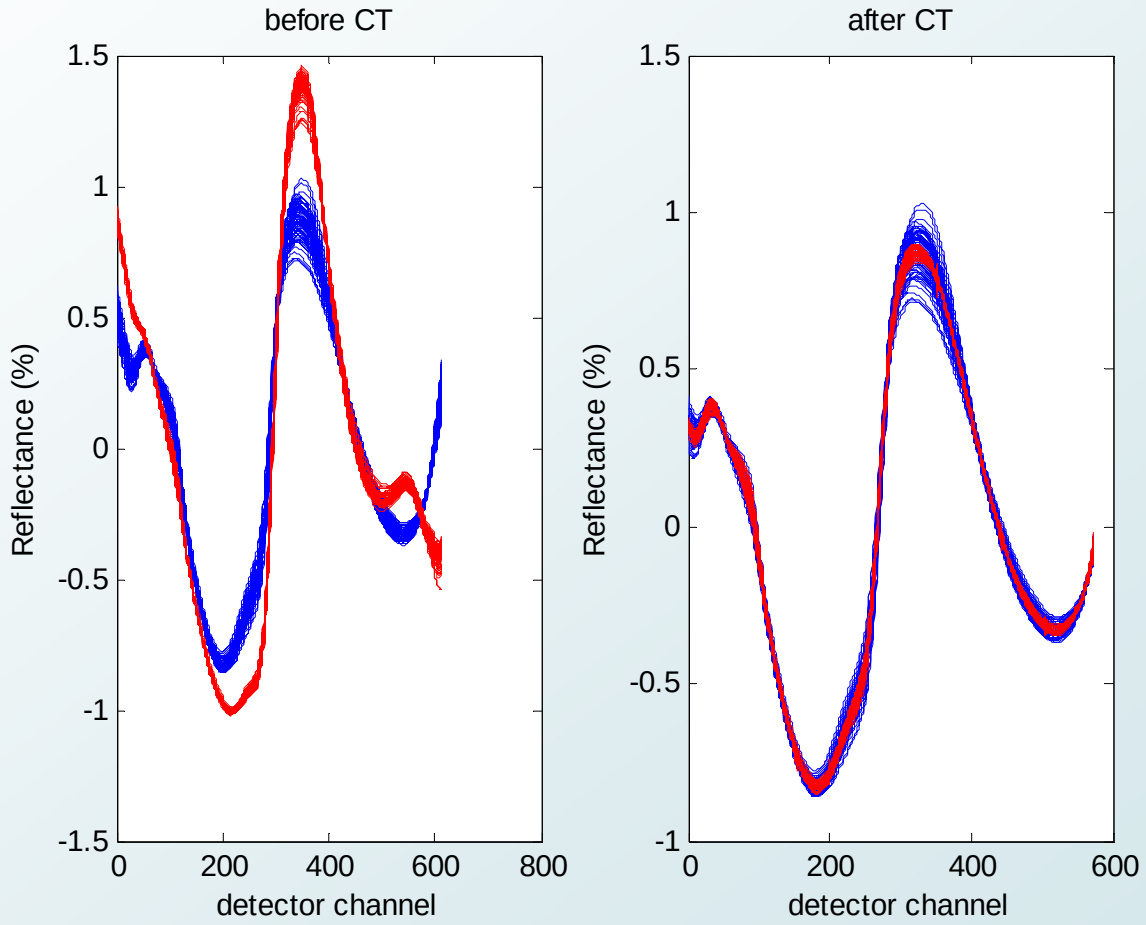




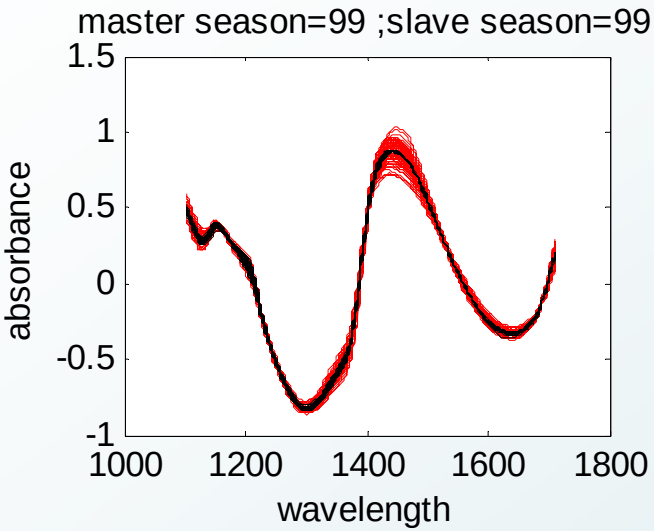


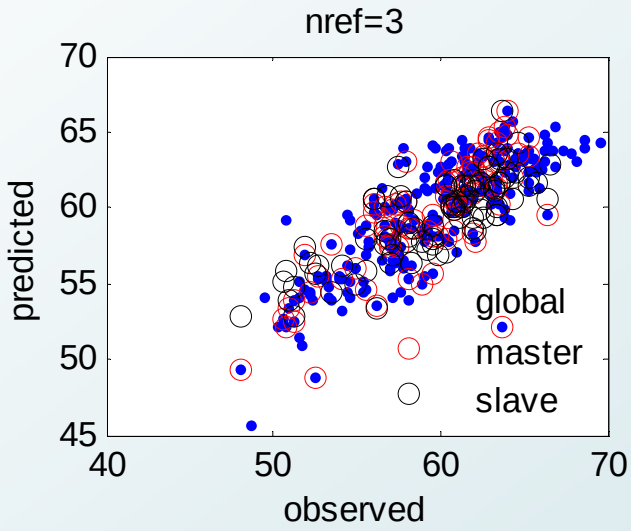


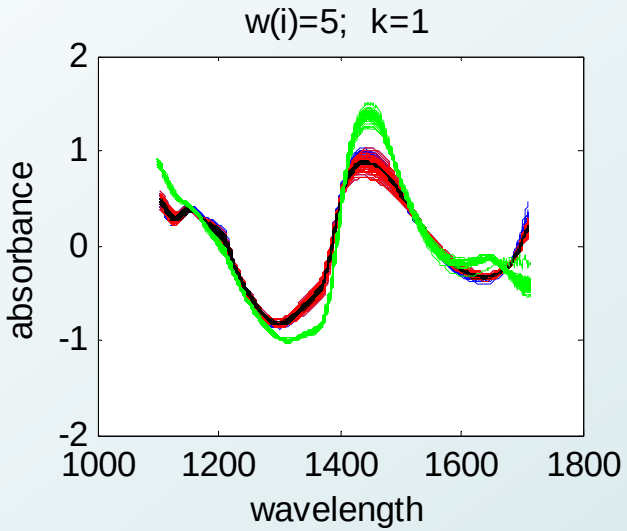
Piecewise Direct Standarization after pre-treatment (MSC Detrend)

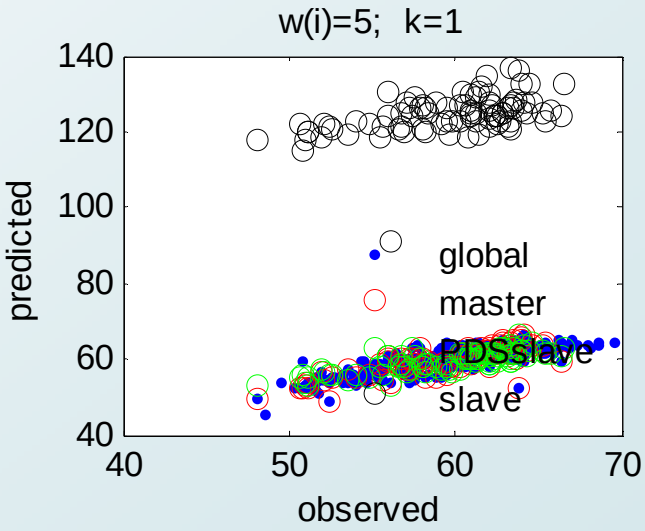


PDS 2D

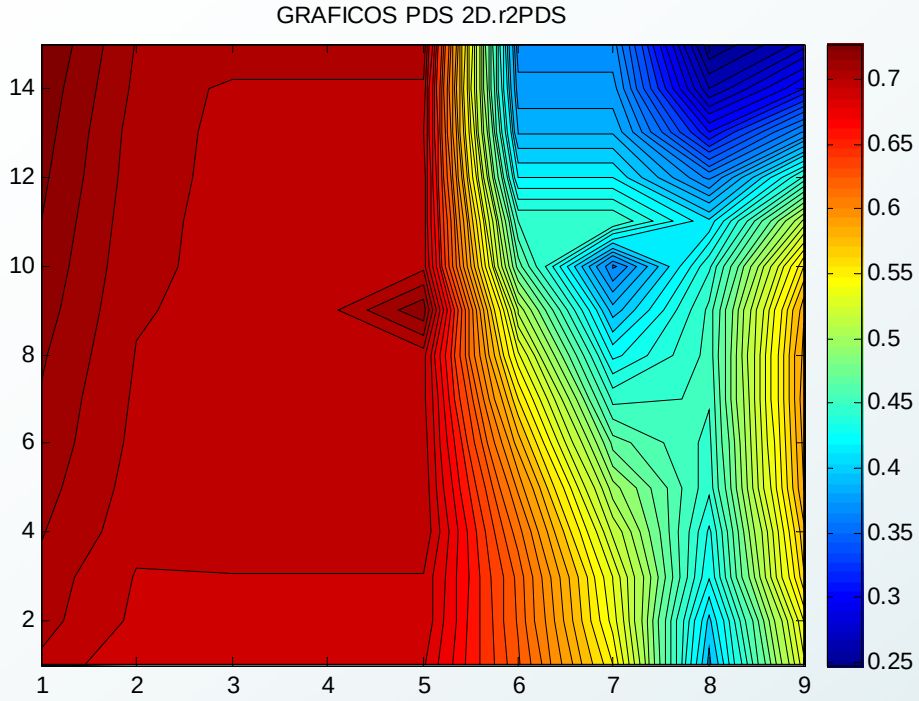

master season=99 ;slave season=99
absorbance
wavelength
1.5
1
0.5
0
-0.5
-1
1000
1200
1400
1600
1800

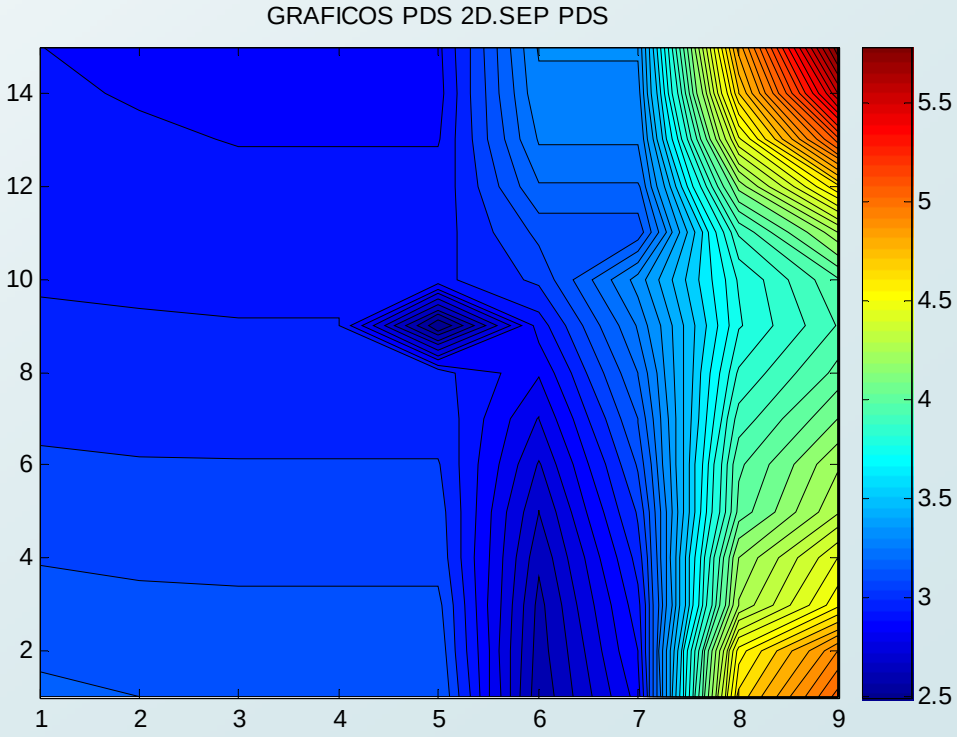

nref=3
predicted
observed
70
65
60
55
50
45
40
50
60
70
global
master
slave


w(i)=5; k=1
absorbance
wavelength
2
1
0
-1
-2
1000
1200
1400
1600
1800


w(i)=5; k=1
predicted
observed
140
120
100
80
60
40
40
50
60
70
global
master
PDSslave
slave

Analysis of optimal solution





Transformation Methods for Dimensionality Reduction

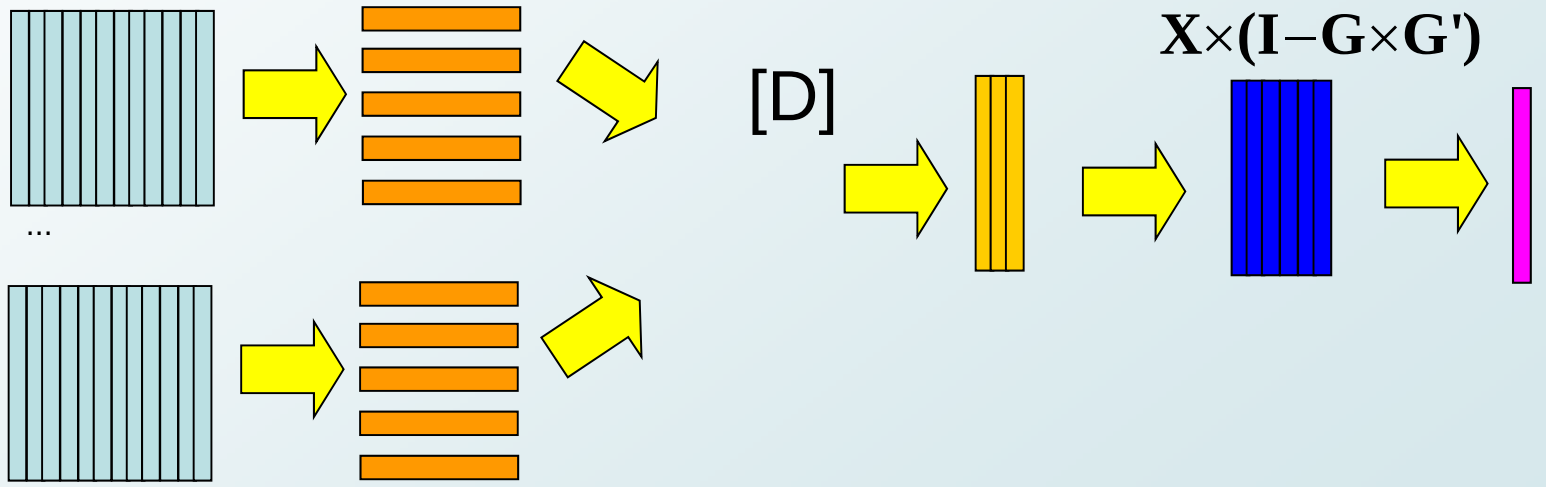
METHOD	DESCRIPTION	REFERENCE
Principal Component Analysis (PCA)	PCA can be used to reduce the data dimensionality, by replacing the original variables by the selected principal components.	(Barros and Rutledge 2004)
Independent component analysis (ICA)	The original data values for each sample may be replaced by the 'scores' or coordinates of that individual on the direction of the 'pure' signal extracted from the mixture of signals in the original data set.	(Zeaiter, Rutledge et al. 2009)
Fourier transform (FT)	The original signal is represented as a sum of sinusoids of different intensities and frequencies.	(Zeaiter, Rutledge et al. 2009)
Wavelet transform (WT)	Recursive application of a matrix of wavelet filter coefficients to a signal, while changing its localization (by translation) and its frequency (by scaling).	(Trygg and Wold 1998)
Simulated annealing (SA)	Variable selection by means of a meta-heuristic algorithm.	(Swierenga, Wülfert et al. 2000)

Orthogonal Projection Methods for Dimensionality Reduction

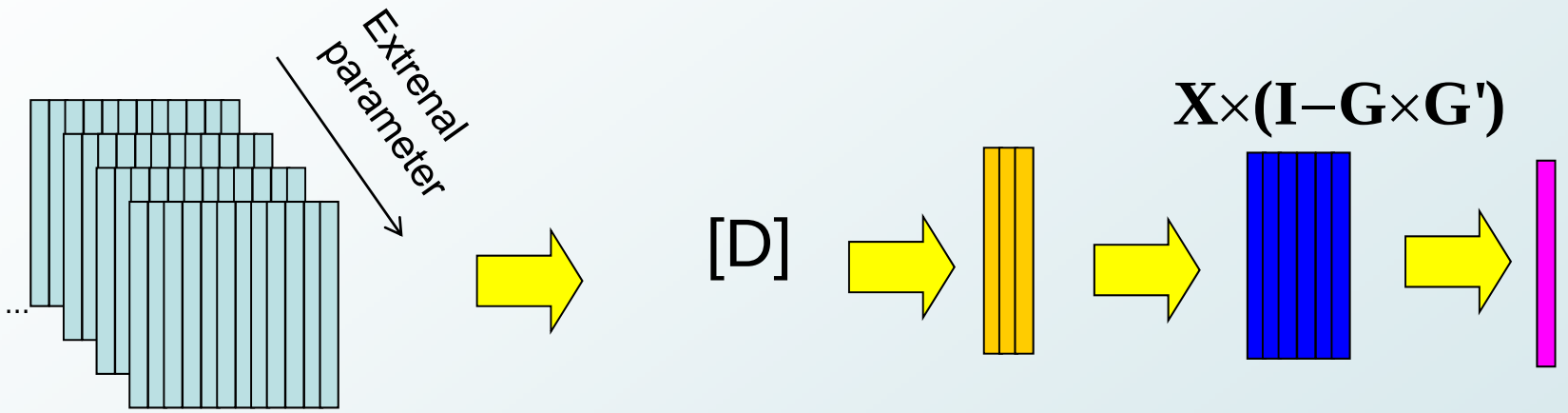
METHOD	DESCRIPTION	REFERENCE
Orthogonal signal correction (OSC)	Determine the corresponding latent structures by PCA, orthogonalized them to y and remove them from the original data	(Wold, Antti et al. 1998)
Projected orthogonal signal correction (POSC)	Intrinsic method. Indirect approach.	(Trygg and Wold 2002)
Direct orthogonal signal correction (DOSC)	Intrinsic method. Indirect approach.	(Westerhuis, de Jong et al. 2001)
Net analyte signal (NAS)	Intrinsic method. Indirect approach.	(Goicoechea and Olivieri 2001)
Ridge-estimated OSC (REOSC)	Improve the performance of the OSC by retaining information related to the analysis	(Shen, Jiang et al. 2006)
Direct orthogonalization (DO)	Intrinsic method. Direct approach.	(Andersson 1999)
Orthogonal projection to latent structures (OPLS)	Intrinsic method. Direct approach.	(Fearn 2000)
Improved piecewise orthogonal signal correction	Intrinsic method. Direct approach.	(Feudale, Tan et al. 2003)
Constrained principal components analysis (CPCA)	Extrinsic method that incorporates external information into the calculation of the PCA of a data matrix	(Takane and Shibayama 1991)
Independent interference reduction (IIR)	Decomposes the data matrix according to the external information (external analysis), and then applies PCA to decomposed matrices (internal analysis).	(Hansen 2001)
External parameter orthogonalization (EPO)	The space that represents the variations due to the perturbations is identified and therefore the spectra are corrected by orthogonal projection.	(Roger, Chauchard et al. 2003)
Dynamic orthogonal projection (DOP)	Virtual standard spectra are created by estimation and then used for orthogonal projection.	(Zeaiter, Roger et al. 2006)

Dynamic orthogonal Projection

[virX,virY]=virtual_spec(TODO,nref,nspec,nwave,tol)
[XXclu,XX]=DPO(Xclu,X,Xclu_r,X_r,ncomp)
TODO_DPO=DPO_allseasons(TODO,nspec,virX,,sel_data,ncomp)



External Parameter Orthogonalization



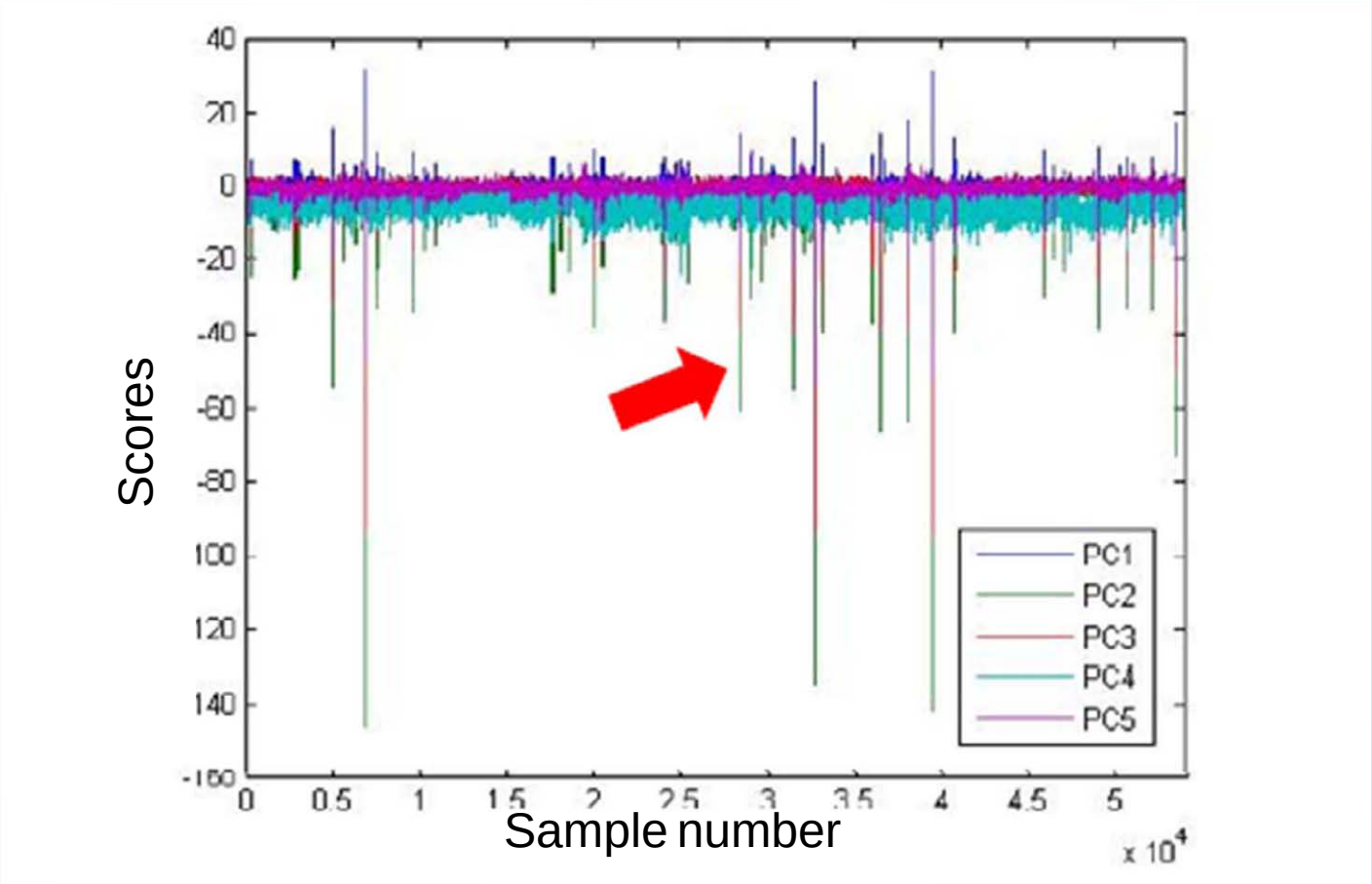
Several attempts of improving robustness

	Wavelength range	spectral variables	*lv*	*g*	r^2	RPD	VD
MLR	894-1284	7	-	-	0.50	1.4	7.349 ppm
PLSR	894-1637	240	12	-	0.64	1.6	100 %
PLS-VSEL	894-1479	26	8	-	0.29	0.9	16.6 %
EPO	894-1637	240	4	3	0.52	1.4	2.256 ppm
DOP 240	894-1637	240	6	3	0.63	1.6	203 ppm
DOP 150	894-1358	150	6	3	0.62	1.6	185 ppm

Multivariate Statistical Process Control Methods

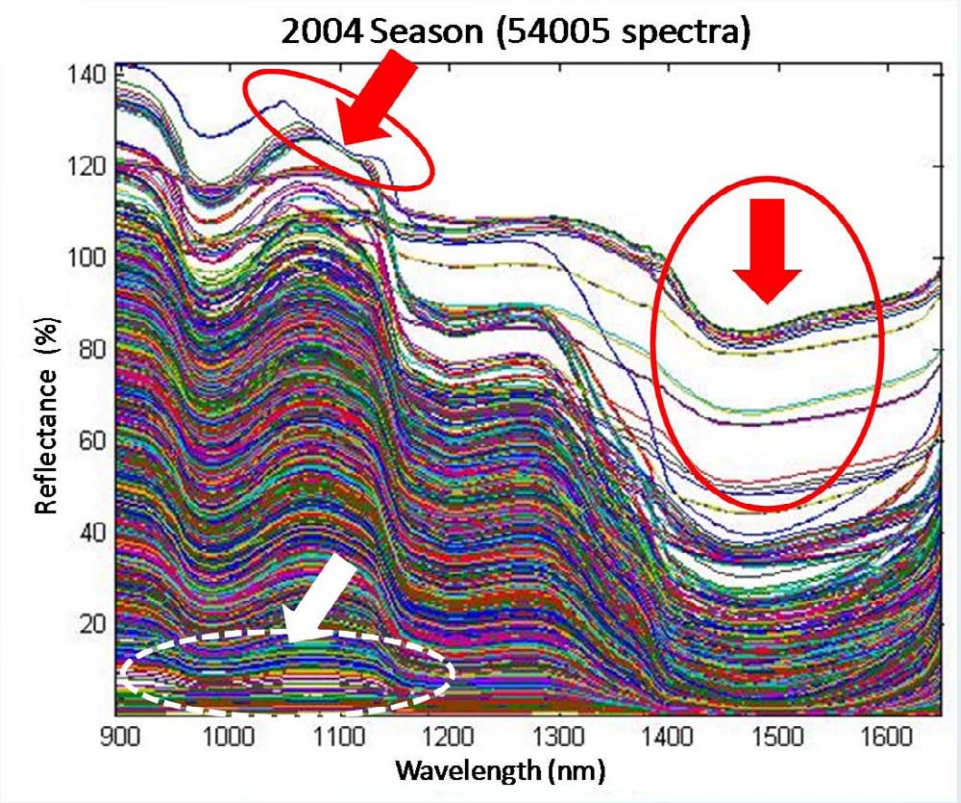
METHOD	DESCRIPTION	REFERENCE
Principal Component Analysis (PCA)	Multivariate projection method that extracts new variables maximizing the variance retained by each new variable .	(MacGregor and Kourti 1995)
Partial Least Squares Regression (PLS)	Multivariate projection method that extracts new variables maximizing the variance retained by each new variable in relation to the variable of interest.	(MacGregor and Kourti 1995)
Real Time PCA	Uses less memory than the classical global PCA.	(Strauss and Prinsloo 2007)
Dynamic Principal Component Analysis (DPCA)	PCA for dynamic systems affected by external variations.	(Ku, Storer et al. 1995)
Independent Component Analysis (ICA)	Multivariate projection method that extracts the pure underlying signals from a set of mixed signals in unknown proportions.	(Kano, Hasebe et al. 2004)
Non linear PCA	Operates in different non-linear scales.	(Choi, Morris et al. 2008)
Canonical Variate Analysis (CVA)	Calculates linear combinations of the 'past' values of the system inputs and/or the outputs that are most highly correlated with linear combinations of the 'future' values of the outputs of the process.	(Simoglou, Martin et al. 2002)

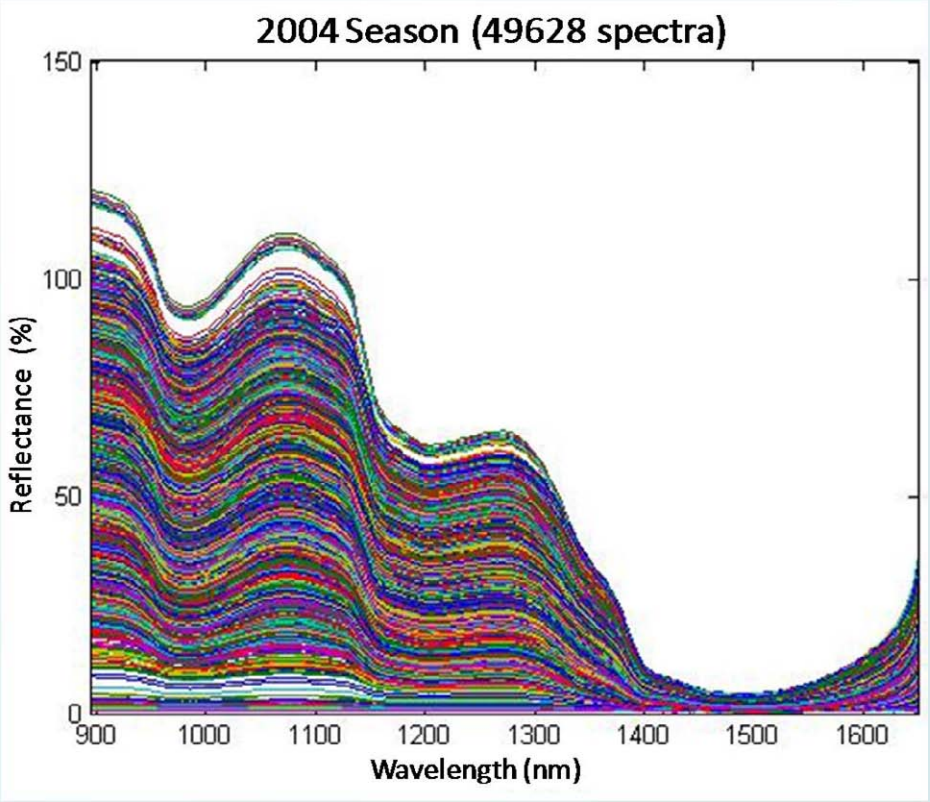
PCA based identification of abnormal spectra



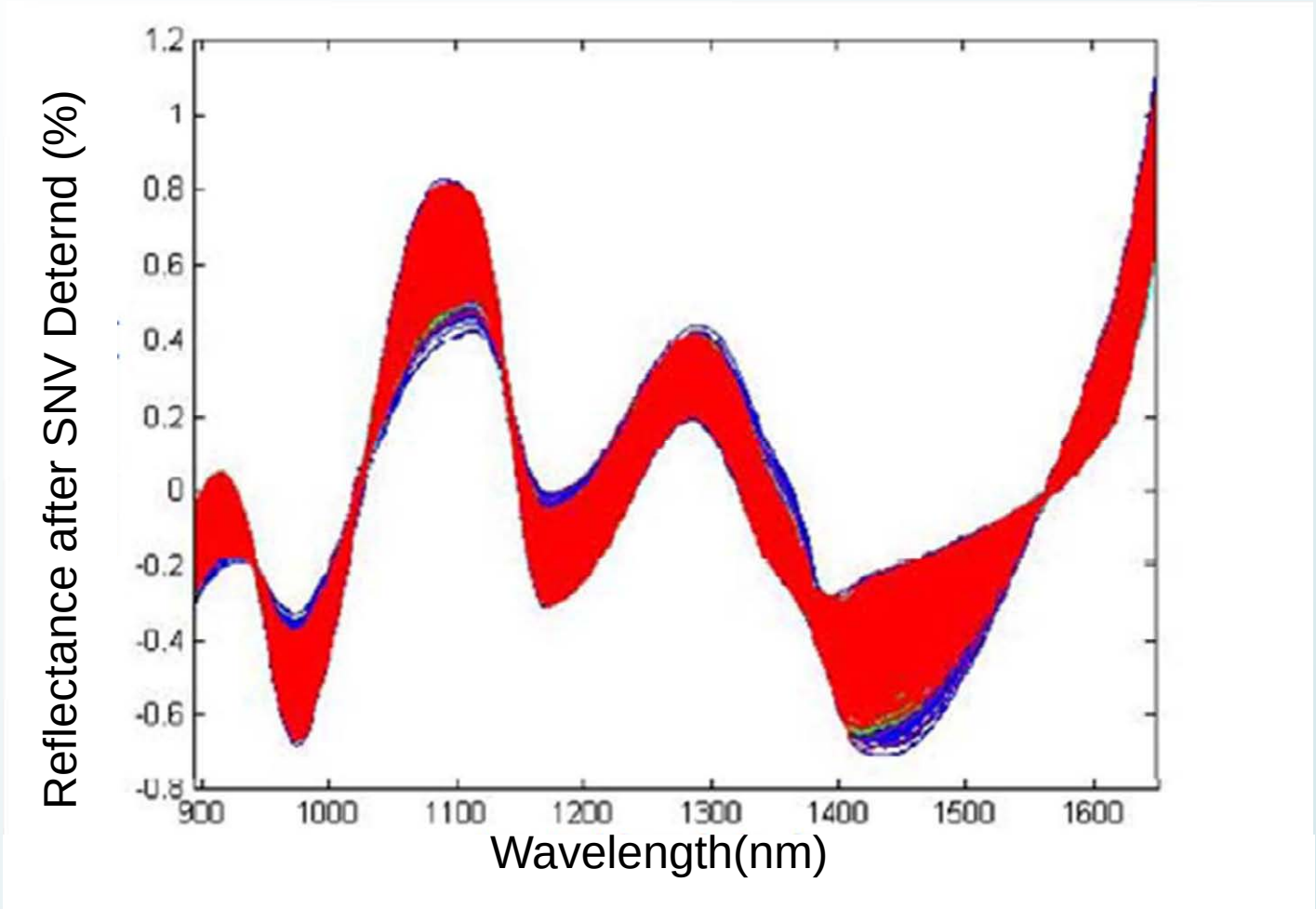
On-line identification of abnormal spectra

PCA based

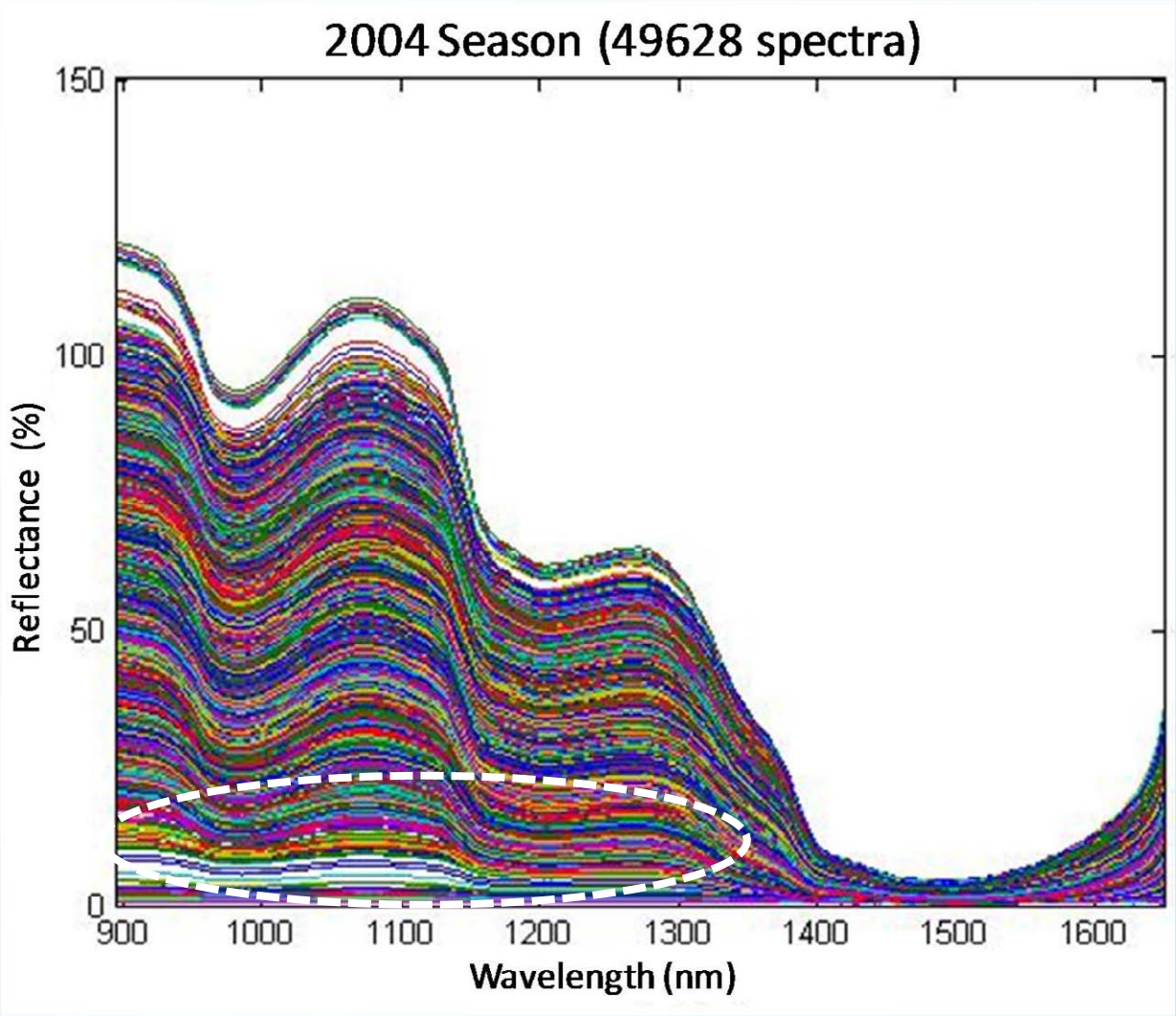




off-line diagnosis



But…


2004 Season (49628 spectra)
Reflectance (%)
150
100
50
0
900
1000
1100
1200
1300
1400
1500
1600
Wavelength (nm)

Multivariate Statistical Process Control (MSPC)

(PCA based)

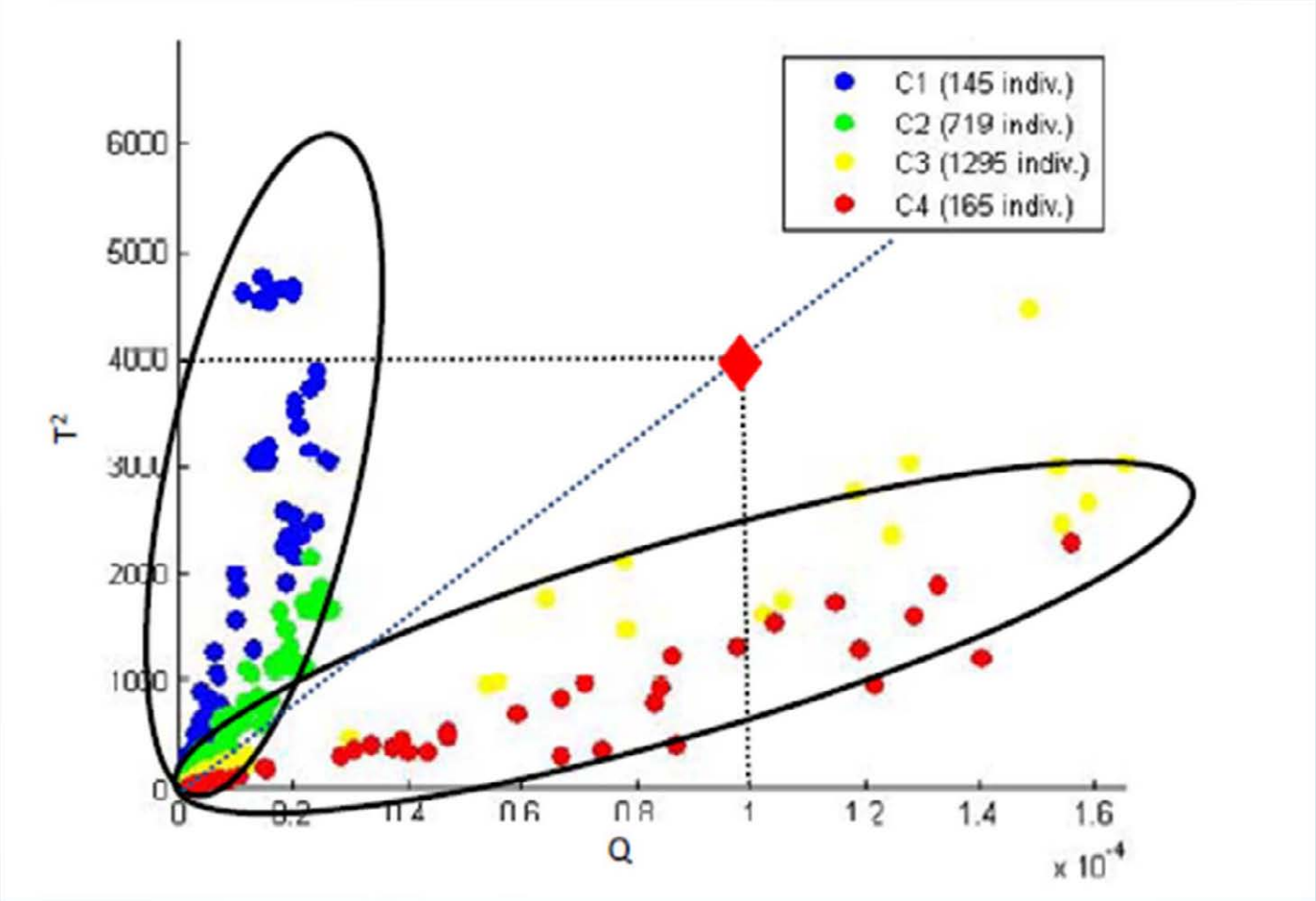
- **Hotelling T^2**

$$T^2 = t_k S^{-1} t_k^T \sim \frac{k(n-1)}{(n-k)} F_{k,n-k}$$

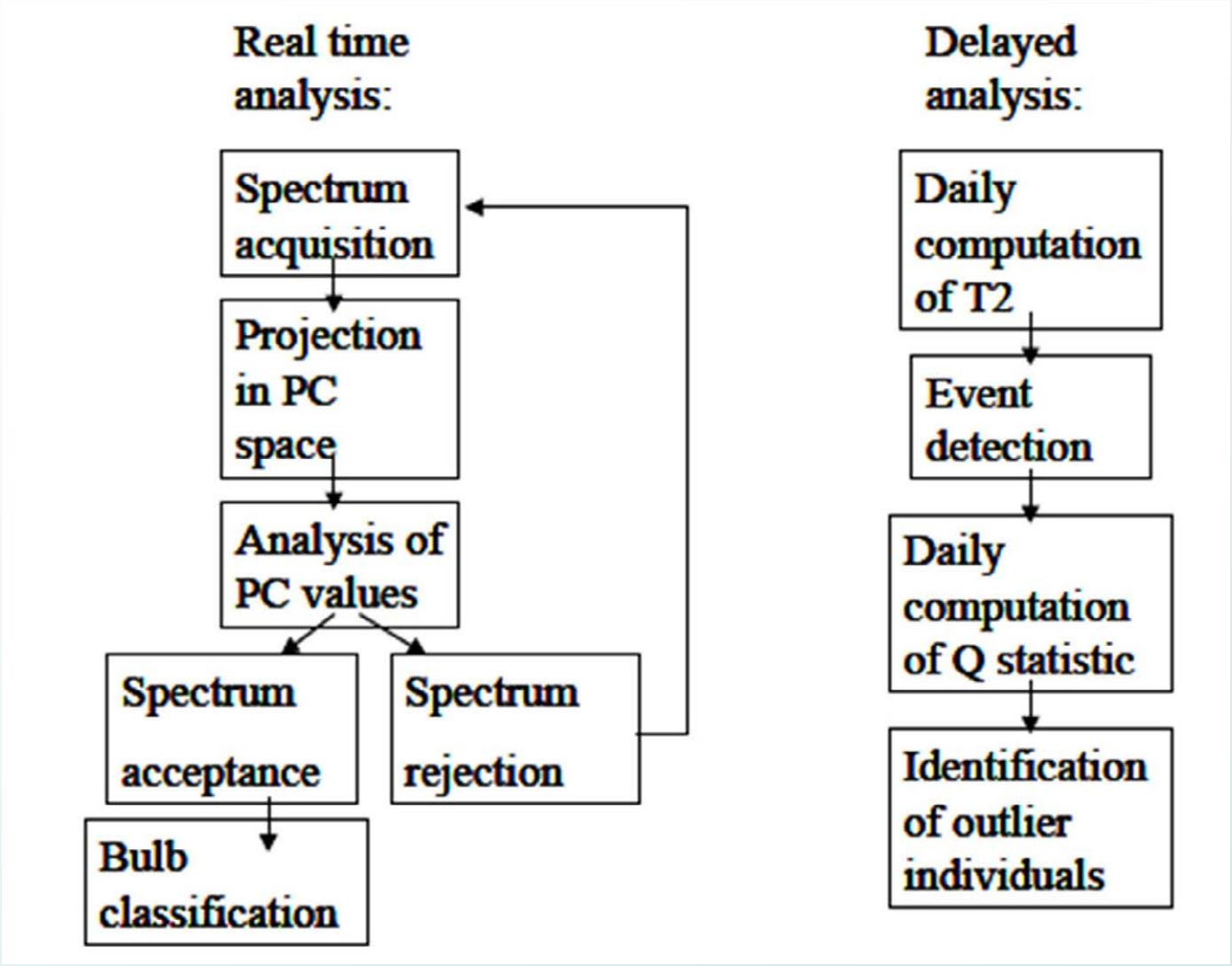
- **Q statistic**

$$Q = \sum_{j=1}^{k} (x_{ij} - \hat{x}_{ij})^2$$

T2-Q Plots


C1 (145 indiv.)
C2 (719 indiv.)
C3 (1295 indiv.)
C4 (165 indiv.)
6000
5000
4000
3000
2000
1000
0
T²
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
Q
x 10⁻⁴

Just to summarize...

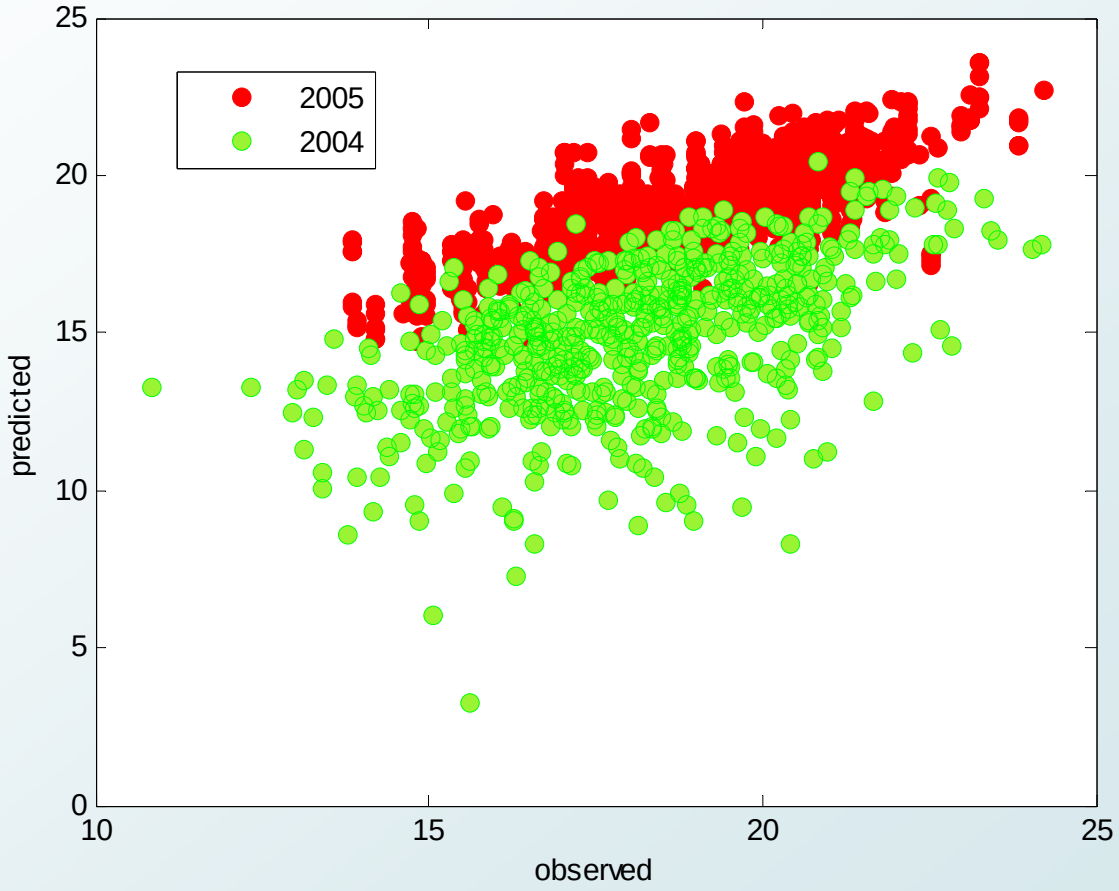


on-line vs off-line control

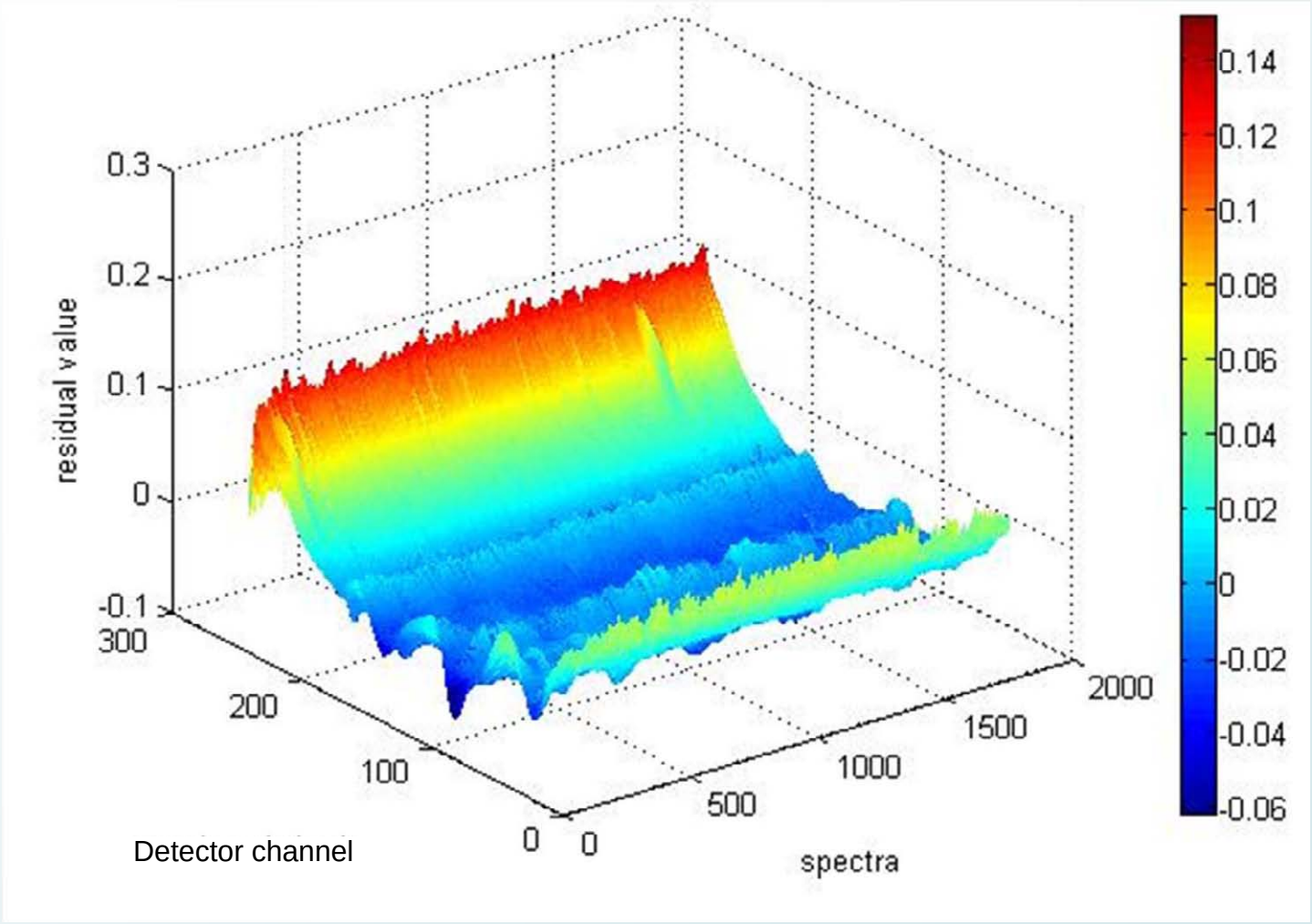
2005 SEASON					
	offline in control	offline out of control			TOTAL
		M1 (T2)	M2(Q)	M3(T2 y Q)	
on line in control	13211	1895	15590	17595	48291
on line out of control	93	60	133	1416	1702
TOTAL	13304	1955	15723	19011	49993
2006 SEASON					
	offline in control	offline out of control			TOTAL
		M1 (T2)	M2(Q)	M3(T2 y Q)	
on line in control	2079	146	9107	14543	25875
on line out of control	119	24	299	4222	4664
TOTAL	2198	170	9406	18765	30539
2007 SEASON					
	offline in control	offline out of control			TOTAL
		M1 (T2)	M2(Q)	M3(T2 y Q)	
on line in control	653	736	15874	43254	60517
on line out of control	11	24	222	4673	4930
TOTAL	664	760	16096	47927	65447
2008 SEASON					
	offline in control	offline out of control			TOTAL
		M1 (T2)	M2(Q)	M3(T2 y Q)	
on line in control	6823	705	19057	35897	62482
on line out of control	172	32	713	5515	6432
TOTAL	6995	737	19770	41412	68914

What to do if PLS is not transferable from one season to another even though spectral treatment

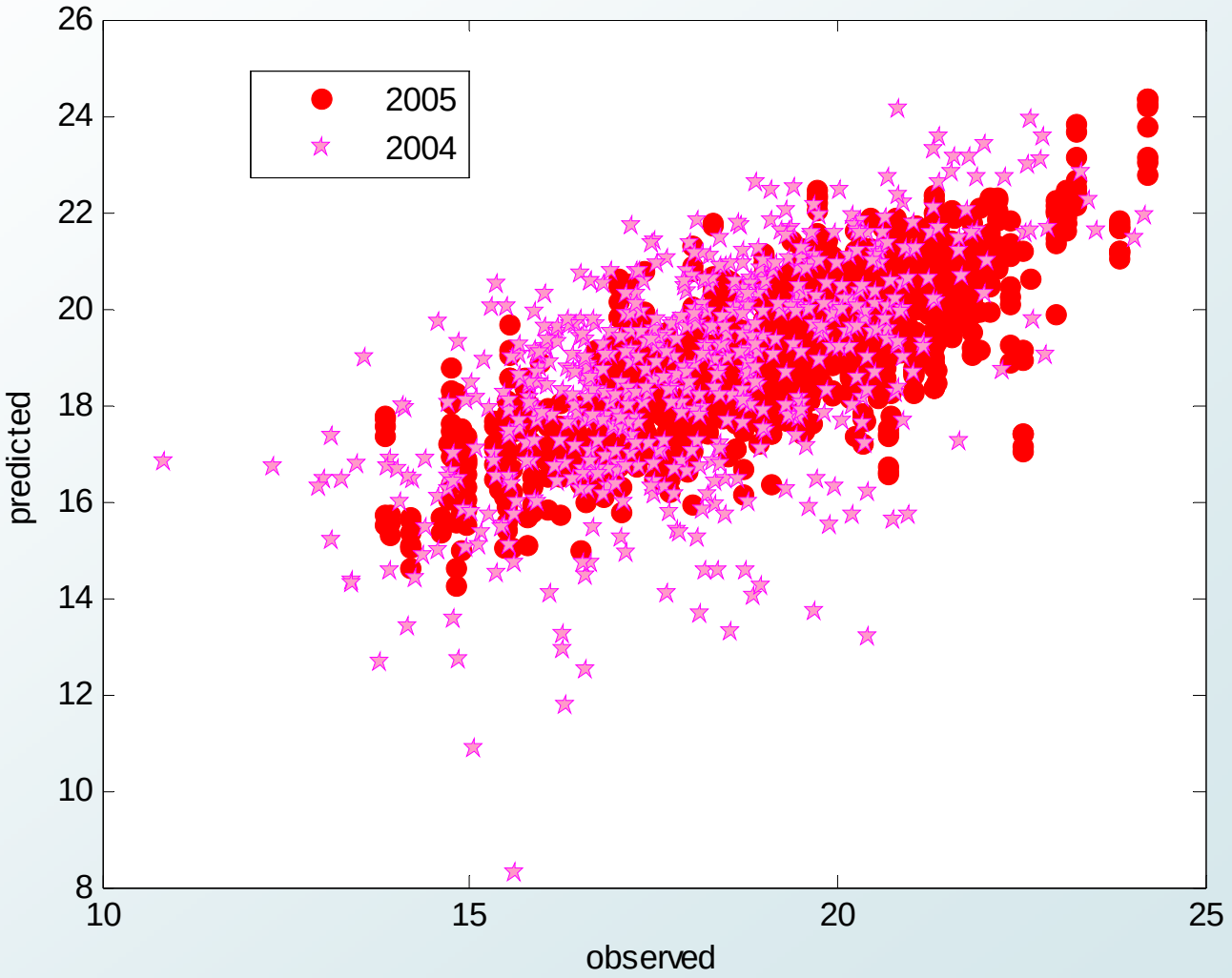
Is it the effect of residuals at certain wavelengths?



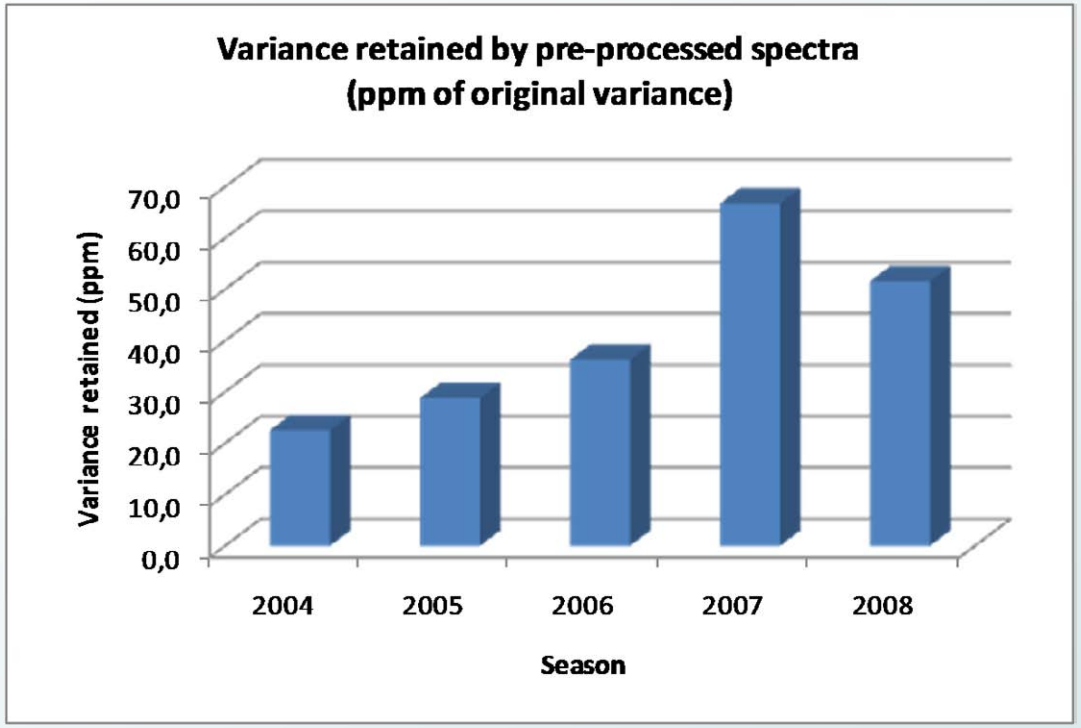
Wavelength prunning before PLS

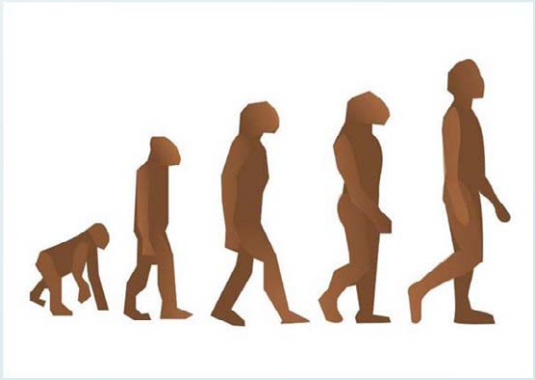


When the CT ends... are you any better?

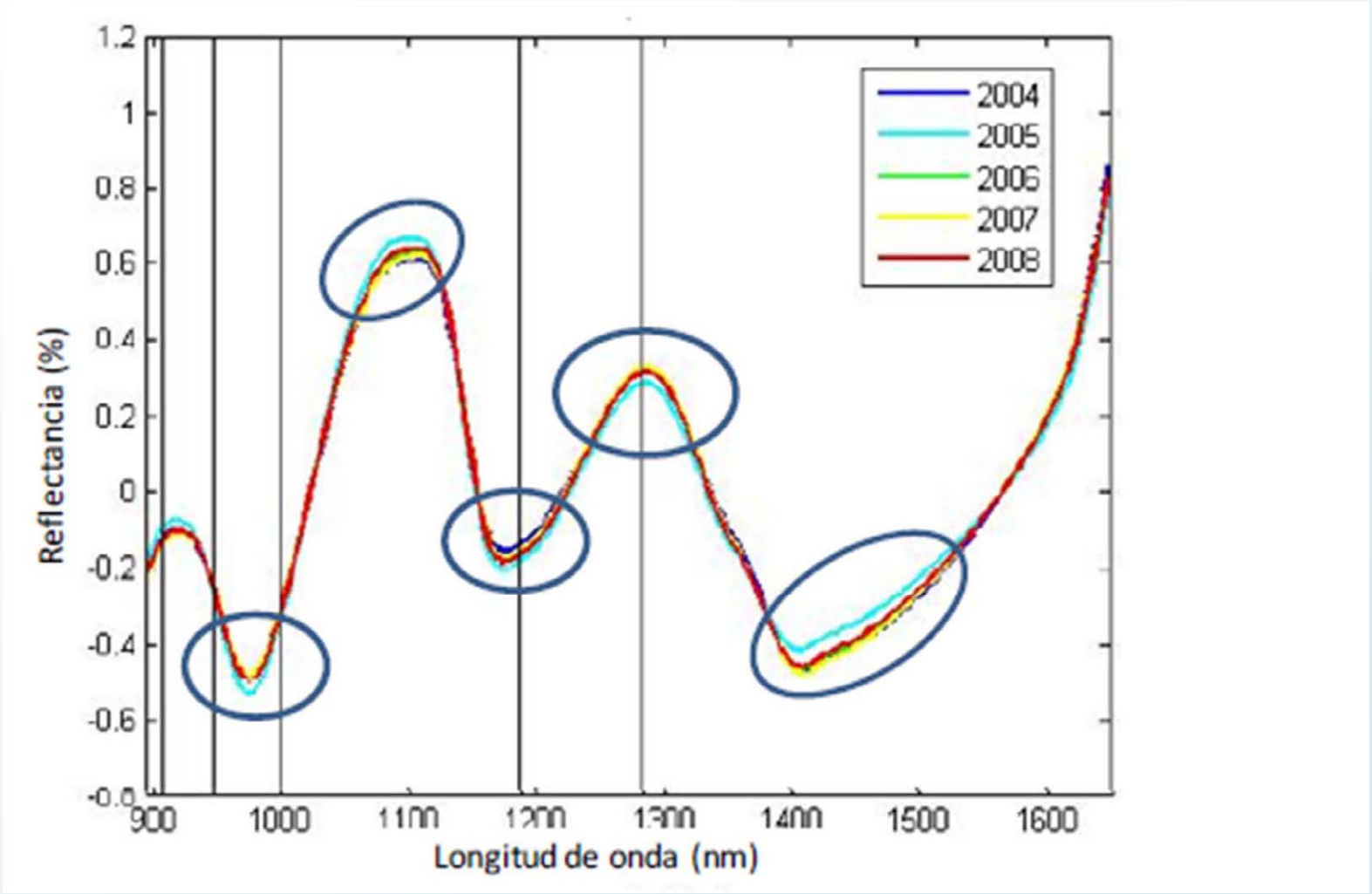


What about the Agroevolution?



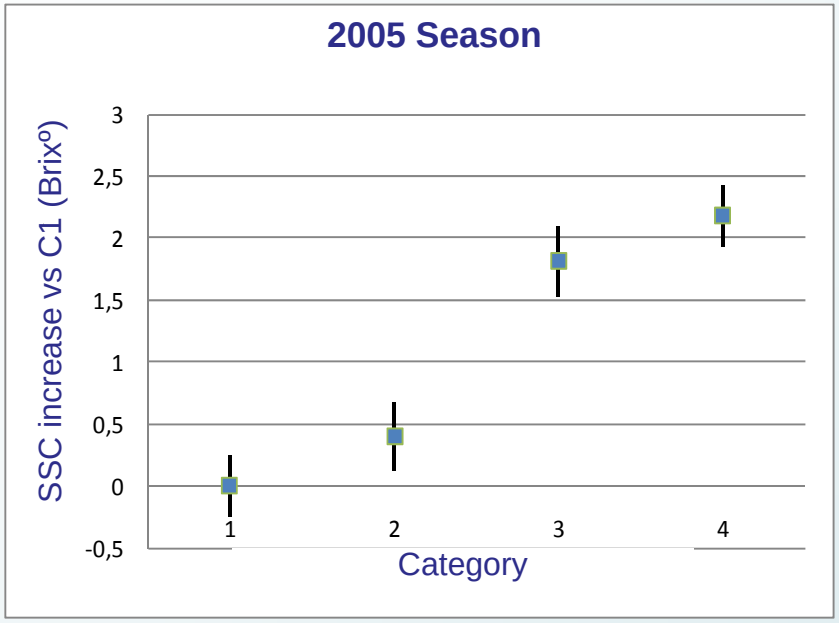


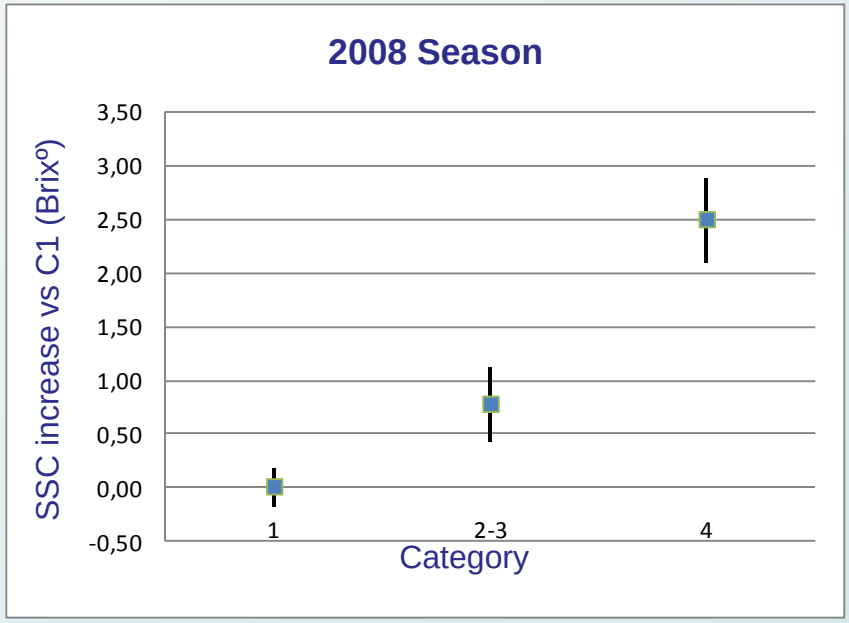
See the changes on season average spectra...



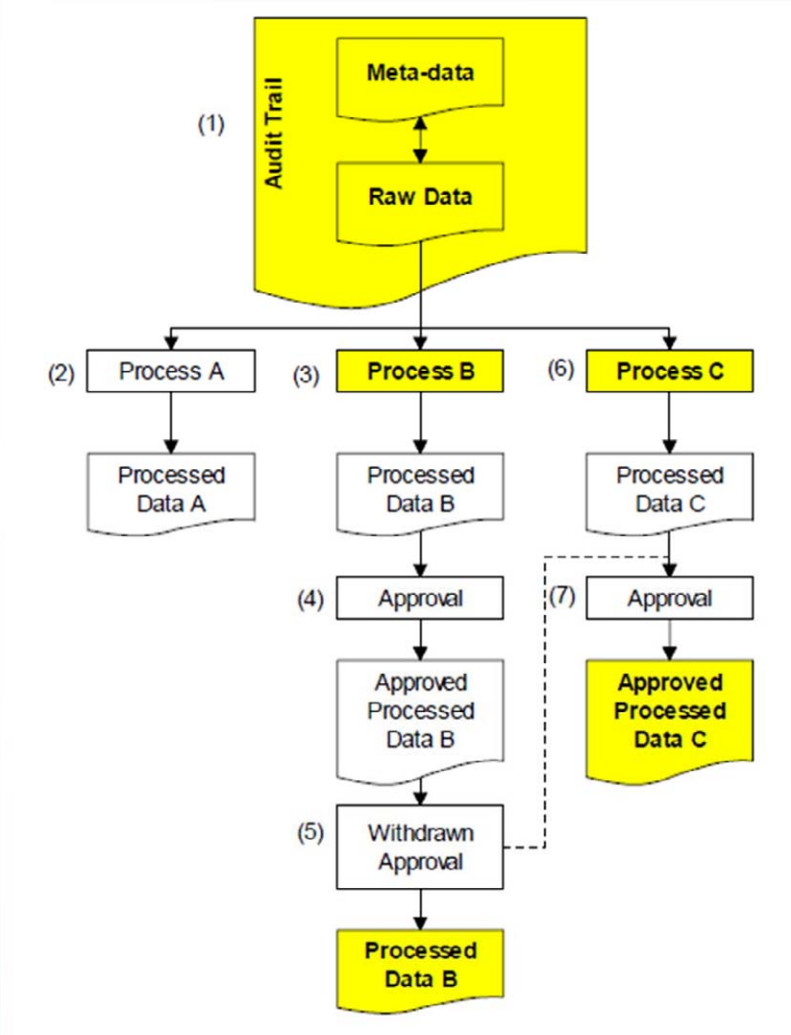
Qualitative classification through seasons

Driscrimination ability





What about Calibration transfer in the context of GLP?



GLP
Working Group on
Information Technology
2005

So the Calibration Transfer Concept is ...

To run all the time, contrawise all other variations, trying to remain in the same position



Having settled the proper level, amount, and variety of mental food, it remains that we should be careful not to swallow the food hastily without mastication, so that it may be thoroughly digested

Feeding the mind

Lewis Carroll

1884



First African-European Conference on Chemometrics

Rabat, Morocco, September 2010